

UNIVERSAL
LIBRARY

OU 160390

UNIVERSAL
LIBRARY

OSMANIA UNIVERSITY LIBRARY

Call No. 313.4 * / B896 Accession No

Author Perumal, S. S.

Title Religion in India.

This book should be returned on or before the date last marked below.

BUSINESS STATISTICS

by

MARTIN A. BRUMBAUGH, Ph.D.

Professor of Statistics, University of Buffalo

and

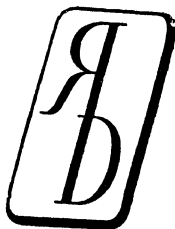
LESTER S. KELLOGG, M.A.

*Assistant Professor of Statistical Research
Ohio State University*

with the collaboration of

IRENE J. GRAHAM, M.A.

*Laboratory Assistant, Department of Statistics
University of Buffalo*



1950

RICHARD D. IRWIN · INC
CHICAGO

**COPYRIGHT 1941 BY RICHARD D. IRWIN • INC.
ALL RIGHTS RESERVED • THIS BOOK OR ANY PART THEREOF MAY NOT
BE REPRODUCED WITHOUT THE PERMISSION OF THE PUBLISHER**

FIRST EDITION

First PrintingSeptember, 1941
Second Printing.....August, 1942
Third Printing.....February, 1946
Fourth Printing.....August, 1946
Fifth Printing November, 1946
Sixth PrintingMay, 1947
Seventh Printing October, 1947
Eighth PrintingFebruary, 1949
Ninth PrintingSeptember, 1949
Tenth Printing October, 1950

PREFACE

AS STATISTICAL METHODS have been gradually expanded in recent years, textbook writers on the subject have exhibited a noticeable tendency to increase the amount of advanced material at the expense of the elementary content. The authors of this book hold to the opinion that the first requisite of an adequate structure is a sound foundation. Pursuant to this point of view they have attempted to place unusual emphasis upon the elementary or foundation methods of the subject. No attempt has been made to attain the research frontiers of any phase of statistical analysis. In short the aim is to present statistical materials and methods that are in everyday use in the conduct of business affairs.

The readers of statistical texts might be divided into two broad groups: those who will compile, analyze, and interpret statistical data and those who will be the users of the results prepared by the first group. The latter readers comprise much the larger group and for their use either as students or business practitioners this book provides the essentials of method and the mental conditioning so necessary for effective "statistical consumption." The training requirements of the first group, the statistical producers, differ somewhat according to the level at which they engage in statistical work. Those few who are conducting advanced research work have little need ~~for~~ this book. The larger number who either now or in the future contemplate engaging in the usual type of statistical collection, analysis, and presentation carried on within business concerns, statistical organizations, and governmental agencies will find that the contents of this book provide sound guidance.

The student will quickly discover that the methods of statistics form a related whole; that the division into chapters is for convenience in the classroom rather than for the separation of disconnected subject matter. The structure is built literally method upon method like the bricks in a wall and, to carry the analogy a step farther, the binding mortar is reasoning rather than memory. The student who attempts to acquire statistical knowledge solely by memorizing rules and formulas invariably fails to develop the power to apply his skill to practical problems. On the other hand the student who approaches the

subject with an eternal "Why?" and insists on having his curiosity satisfied has a much better chance of developing the power needed to solve new problems as they arise.

This distinction of attitude is exemplified by a criticism often leveled at the writers of statistical texts: "If they would only tell us what techniques to use in analyzing different kinds of data they could save so much time and take most of the mystery out of the subject." If the quoted suggestion could be put into effect, statistical practice would in truth be reduced to a routine level. But the case is not so simple because the type of analysis that should be applied to a particular set of data depends entirely on the purpose of the analysis, the specific use that will be made of the results, the time and funds that are available, and other related considerations. Therefore neither in this book nor in any other will rules be found relating methods of analysis to data on specific subjects. The function of the statistician must always be the exercise of judgment as to the proper methods to employ in the investigation of a given problem.

From the standpoint of mathematics, this book assumes only that the reader has proficiency in arithmetic and enough knowledge of algebraic symbolism to be able to substitute values in a formula. Even this modest assumption is partially reinforced by the introduction early in the book of a chapter entitled "The Use of Numbers," the content of which is partially a review of arithmetic. The explanations in the book assume that the reader possesses some familiarity with economics and understands in an elementary way the organization and functioning of individual business concerns. A knowledge of accounting principles, marketing principles, and recent business history including the relation of government to business provides a desirable although less essential background.

The subject matter of this book can be covered in a ninety-hour course. By reducing the intensity of coverage somewhat, the entire content can be included in a sixty-hour course. For briefer courses some chapters will undoubtedly have to be omitted entirely or in part. It is difficult to specify which chapters might be omitted, since each case requires a knowledge of the point of view of the instructor, the capabilities of the particular group of students, and the purpose of the statistics course in the curriculum.

The problems appearing at the end of each chapter have been so planned that the student who prepares answers to all or a large part

of them will be forced to apply the important principles of the subject to situations akin to those found in statistical practice. The authors hold steadfastly to the opinion that effective teaching of statistics requires the liberal use of problems to insure facility in computation, to provide constant practice in selecting and adapting methods to particular situations, to develop proficiency in the interpretation of the results of investigation, and to encourage accurate reporting of completed analyses. Many of the problems have been reduced in size to accommodate them to the needs of students, but an attempt has been made to avoid simplification to the point of absurdity.

A list of references is appended to each chapter. These lists are intended to be selective rather than comprehensive. The publications named are included because they contain material which supplements the development in the text or because they offer the opportunity for either more intensive or more advanced study of the subject matter in the text. The reference lists are not intended to set forth the sources from which assistance was drawn for this book. Wherever such assistance is specific, direct footnote reference has been made to the source and the author's permission for use has been secured. For aid in the broader sense the authors are permanently indebted to writers in the field of statistics in so many directions that anything beyond general acknowledgment of the obligation would be impossible. If, by some mischance, the authors have failed to acknowledge materials reproduced from other writers, such omission is wholly unintentional.

It is impossible to make a complete acknowledgment to Miss Irene Graham, who has collaborated with the authors in the preparation of this text. Her most outstanding work was the writing of chapters XIII and XIV and major revision of chapter XI. She has also contributed materially to the text of chapters IV to X, XV, XVII, XIX, and XXX, in addition to research, criticism, reorganization and editing of all chapters.

Dr. Robert Riegel, professor of statistics, University of Buffalo, has read and criticized the manuscript in various stages of preparation. His suggestions concerning statistical soundness and pedagogical desirability represent a contribution that the authors acknowledge gratefully.

Chapter XXVI has been made possible through the co-operation of Mr. A. H. Robinson, assistant-treasurer, and Mr. Lawrence M. Tarnow, head of the planning department of Eastman Kodak Company of Rochester, New York. To them for their assistance, and to the Eastman

Kodak Company for permission to use the material, the authors express their sincere thanks.

Most of the graphs were prepared by Mr. Ralph Lownie, a student in the School of Business Administration, University of Buffalo. Their quality is due entirely to his craftsmanship. The remaining graphs were drawn by Mrs. Dorothy Tallman, who also shared with Mrs. Ruth Carroll the extremely laborious task of typing the manuscript.

M. A. BRUMBAUGH

LESTER S. KELLOGG

September, 1941

TABLE OF CONTENTS

CHAPTER	PAGE
I. STATISTICS IN BUSINESS	1
The Statistical Approach. The Work of the Statistician.	
II. THE USE OF NUMBERS	13
Introduction. The Fundamental Operations. Fractions. Square Roots.	
Accuracy of Statistical Data. Summary.	
III. STATISTICAL INVESTIGATION	43
The Character of Statistical Investigation. The Canons of Statistical	
Investigation. Steps in Statistical Investigation. The Scope of Dif-	
ferent Investigations. Summary.	
IV. PRELIMINARY PLANNING OF INVESTIGATIONS	53
Introduction. Define the Problem. Study the Problem. Plan the	
Procedure. Prepare a Statement of the Program.	
V. SAMPLING	69
Relation to Knowledge. The Importance of Sampling. The Prin-	
ciple of Statistical Regularity. The Two Problems of Sampling.	
VI. COLLECTION OF DATA—DIRECT SOURCES	92
Description of Direct Sources. Collecting Data From Direct Sources.	
Summary.	
VII. EDITING AND PRELIMINARY TABULATION	118
Editing Schedules. Preliminary Tabulation.	
VIII. TABULATION	144
Definitions. Types of Tables. Established Practice in the Construc-	
tion of Tables. Tabular Forms.	
IX. CLASSIFICATION OF LIBRARY SOURCES	186
The Meaning of Collection From Library Sources. Methods of	
Classifying Sources. Appendix A—Selected Sources Listed According	
to Publishing Agency, Title, Frequency of Publication, and Contents.	
X. THE USE OF LIBRARY SOURCES	210
Introduction. Finding a Good Source. The Correct Use of Data.	
Appendix B—Examples of Search for Data in Libraries.	
XI. RATIOS	229
The Importance of Ratios in Statistics. Construction of Statistical	
Ratios. Presentation of Ratios. Comparisons Between Ratios.	

CHAPTER	PAGE
XII. APPLICATIONS OF RATIOS	274
Refined Ratios. Compound Ratios. Examples of the Use of Ratios in Business.	
XIII. GRAPHS	298
Introduction. Simple Types of Graphs—Methods and Purposes of Each. Introduction to Two-Dimensional Linear Graphs.	
XIV. GRAPHS (continued)	323
Time Series Graphs. Planning Graphs for General Effect.	
XV. FREQUENCY DISTRIBUTIONS AND GRAPHS	350
Frequency Distributions. Graphs of Frequency Distributions.	
XVI. MEASURES OF CENTRAL TENDENCY—AVERAGES OF CALCULATION	387
Introduction. The Arithmetic Average. The Geometric Average.	
XVII. MEASURES OF CENTRAL TENDENCY—AVERAGES OF POSITION	413
The Median. The Mode. Criteria for Selecting and Judging Averages.	
XVIII. MEASURES OF DISPERSION AND SKEWNESS	436
Introduction. Dispersion. Skewness. Uses of Measures of Dispersion and Skewness.	
XIX. INDEX NUMBERS	461
Introduction. Kinds of Index Numbers. Basic Methods of Constructing Index Numbers. Symbols. Unweighted Formulas. Weighted Formulas. Problems of Index Number Construction. Tests of Index Numbers. Specific Uses of Index Numbers and Their Interpretation.	
XX. SOME COMMONLY USED INDEXES	506
Introduction. An Index of Cost of Living. An Index of Industrial Production. An Index of Employment. A Wholesale Price Index. A Local Business Index.	
XXI. ANALYSIS OF TIME SERIES	537
A Change in Emphasis. Components of Time Series. The Problem of Time Series Analysis. Preliminary Analysis.	
XXII. TREND	560
The Location of Trend. Methods of Measuring Trend. Why Trend is Measured. Summary.	
XXIII. SEASONAL AND CYCLICAL MOVEMENTS	592
Introduction. The Nature of Seasonal Variation. The Concept of a Seasonal Pattern. Methods of Measuring Seasonal Variation. The Cyclical Remainders. Special Seasonal Problems.	

TABLE OF CONTENTS

ix

CHAPTER	PAGE
XXIV. SUMMARY OF THE ANALYSIS OF TIME SERIES (An Example)	627
Adjustment of Calendar Variation. Adjustment of Changes in the Price Level. Adjustment of Seasonal Variation. Adjustment of Trend—Moving-Trend Method. The Cyclical Fluctuations.	
XXV. INDEXES OF BUSINESS CONDITIONS	649
Need for External Information. Two Types of Business Indicators. Construction of Composite Indexes. Examples of the Construction of Composite Indexes. Local Indexes. Use of Business Indexes.	
XXVI. INTERNAL APPLICATION OF TIME-SERIES ANALYSIS PRODUCTION PLANNING	682
Introduction. General Discussion of Production Planning. Two Examples of Planning. Summary.	
XXVII. CORRELATION	704
Introduction. Scattergram. The Regression Line. The Standard Error of Estimate. The Coefficient of Correlation. Some Deferred Points. The Rank Difference Measure of Correlation. Correlation of Time Series. Recapitulation of Formulas.	
XXVIII. THE NORMAL CURVE	752
Probability. Binomial Distribution. The Normal Curve. Other Types of Distributions. Goodness of Fit—Chi-Square Test.	
XXIX. PRINCIPLES OF SAMPLING AND TESTS OF SIGNIFICANCE . .	785
Introduction. The Basis of Sampling. Tests of Significance. Small Samples. Variance Analysis.	
XXX. PRESENTATION OF THE RESULTS OF STATISTICAL INVESTIGATION	826
Introduction. The Writer—Reader Relation. Requirements of a Report. The Form of a Report.	
APPENDIX A—SELECTED SOURCES LISTED ACCORDING TO PUBLISHING AGENCY, TITLE, FREQUENCY OF PUBLICATION, AND CONTENTS	197
APPENDIX B—EXAMPLES OF SEARCH FOR DATA IN LIBRARIES	224
APPENDIX C—LOGARITHMS OF NUMBERS	845
FIVE PLACE TABLE OF LOGARITHMS	857
APPENDIX D—TABLE OF SQUARES, SQUARE ROOTS AND RECIPROCALs	875
APPENDIX E—TABLE OF ORDINATES	895
APPENDIX F—TABLE OF AREAS	896

FIGURES

FIGURE

PAGE

1. Two Examples of the Development by Successive Steps of the Solution for Extracting Square Root	24-25
2. Schedule Used in a Real Estate Survey	95
3. Proposed Revision of Real Estate Schedule, Figure 2	96
4. Questionnaire Used in Worsted Spinning Spindle Inventory	101
5. Radio Section of Questionnaire Used in Surveying the College Market	105
6. Letter with Appeal for Reply Based on Co-operation	106
7. Letter with Appeal for Reply Based on Interest	107
8. Letter with Appeal for Reply Based on Profit	107
9. Letter with Appeal for Reply Based on Obligation	108
10. Letter with Appeal for Reply Based on Position	108
11. Letter Based on Compulsion	109
12. Example of Instructions to Collecting Agents	110
13. Collection Card Used in Residential Vacancy Investigation in Buffalo, New York	123
14. Tally Sheet for Recording Residential Vacancy in Buffalo, New York	126
15. Proposed Work Sheet for Questionnaire Used in College Market Investigation, Figure 5	128
16. Punched Cards for Mechanical Tabulation	132
17. Schedule Used in Collecting Data for the President's Conference on Home Building and Home Ownership, 1930	134
18. Instructions for Coding Data on Home Building and Home Ownership in Buffalo, New York, 1930	135
19. Code Sheet Used in Transferring Information from Questionnaire, Figure 17	138
20. Reproduction of the Printed Record from the Tabulating Machine with Headings Added	140
21. Form for Two-Way Cross-Classification	157
22. Form for Three-Way Cross-Classification	157
23. Form for Four-Way Cross-Classification	158
24. Form for Five-Way Cross-Classification	159
25. Reproduction of Department of Agriculture Form C. E. 1-128	171
26. Reproduction of Department of Agriculture Long-Term Blank	172
27. Reproduction of Department of Agriculture Form C. E. 1-139	173

FIGURE	PAGE
28. Eastman Kodak Co. Form—Comparison of Sales by Divisions . . .	175
29. Eastman Kodak Co. Form—Lost Time Report	176-177
30. Instructions on the Reverse Side of Lost Time Report . . .	178-179
31. Eastman Kodak Co. Form—Labor Turnover Report	181
32. Organization Chart of the Government of the United States between 192-193	
33. Classification of Comparisons Between Ratios of Like Items with Examples of Each	256
34. Classification of Comparisons Between Ratios of Unlike Items with Examples of Each	264
35. Types of Graphs	299
36. Dot Maps: Filling Stations in the United States, 1935	305
37. Cross-Hatched Ratio Map: Filling Stations per 10,000 Persons in the United States, 1935	308
38. Flow Map: United States Exports, 1931	311
39. Dial Chart: Index of Industrial Activity as of May 31, 1941 . .	313
40. Pictogram: Number of Workers in Basic Fields of Employment, 1940	314
41. Types of Bar Graphs	316
42. Bar Graph of Time Series	323
43. Band Graphs of Time Series: Per Cents and Dollar Values . .	325
44. Line Graphs of Time Series	328-329
45. Construction of the Ratio Scale	333
46. Curves Showing Changing Relative Rates on a Ratio Scale . .	336
47. Types of Lines	341
48. Methods of Plotting Time Periods	345
49. Tally of Monthly Rents Paid by 155 Families in a Consumer Survey in Columbus, Ohio	353
50. Array of Rents Paid by 155 Families in Columbus, Ohio . . .	354
51. Methods of Designating Class Limits	361
52. Two Types of Frequency Diagram of Rent Data	367
53. Frequency Diagrams of Discrete Data. Number of Dresses Sold in Junior Sizes	370
54. Ogives: Cumulative Frequency Diagram of Rent Data	373
55. Frequency Diagrams of Hourly Wage Rates Paid by Fifty-two Industrial Concerns	376
56. Per Cent Comparison of Two Distributions of Rent Data . . .	378
57. Types of Curves	380
58. Lorenz Curves: Cumulative Per Cents of Stores and Sales, Inde- pendent Retail Grocery Stores in Buffalo, 1929 and 1935 . .	383

FIGURE	PAGE
59. Location of the Median in an Array	414
60. Location of the Median in a Frequency Distribution	416
61. Summary of Characteristics of Measures of Central Tendency	432
62. Guide to the Suitability of Measures of Central Tendency Accord- ing to the Condition of the Data	432
63. Comparison of Columbus Rentals with Normal Distribution Accord- ing to Measures of Dispersion	449
64. Fractions of the Area of the Normal Curve Measured by the Standard Deviation	451
65. Summary of Criteria of Measures of Dispersion	454
66. Sources of Commonly Used Index Numbers	465
67. Nomograph for Reading Per Cents of Increase or Decrease in Index Numbers	499
68. National Industrial Conference Board Index of Cost of Living, Monthly, 1923-40	514
69. Federal Reserve Board Index of Industrial Production, Monthly, 1919-40	521
70. Indexes of Employment and Pay Rolls, Monthly, 1919-40	526
71. Wholesale Price Indexes of National Fertilizer Association and United States Bureau of Labor Statistics; Annually, 1929-35; Monthly, 1936-39; Weekly, January 1940-April 1941	531
72. Index of Bank Debits in Canton, Ohio, Monthly, 1926-40	535
73. Production of Wheat, Passenger Automobiles and Anthracite Coal in the United States, 1900-1937, and Free Hand Trend for Each Series	540
74. Consumption of Raw Cotton in the United States 1913-37	543
75. Monthly Sales of F. W. Woolworth Company, 1930-37	546
76. Daily Net Currency Movement in New York City to or from the Federal Reserve Bank of New York, April to September, 1926	548
77. Monthly Totals and Daily Averages for 1936 for Three Sets of Data in Buffalo, New York: Sales of a Drug Store, Flour Milled and Bank Clearings	553
78. Raw Cotton Exported by the United States 1913-22	555
79. Investment in Inventory of Swift and Company Packers, 1913-36	557
80. Moving Averages Fitted to Controlled Data Containing Cycle and Straight Line Trend	564
81. Moving Average Fitted to Controlled Data Containing Cycle and Curved Line Trend	565
82. Number of Horsepower of Diesel Engines Installed Annually, 1918-37	568

FIGURE	PAGE
83. Diagram Used to Write the Equation of a Straight Line	572
84. Parabola Trend Fitted to Postal Receipts at Buffalo, New York, 1920-33	579
85. Logarithmic Trend Fitted to Production of Wood Pulp, 1923-37	580
86. Straight Line Trend Fitted to the Number of Lines of Magazine Advertising, 1913-37	584
87. Relative Cycles of Magazine Advertising, 1913-37	585
88. Straight Line Trend Fitted to Electric Power Production, 1919-29	588
89. Daily Average Consumption of Small Cigarettes, Monthly, 1927-36	595
90. Approximate Method: Test for Seasonal Pattern of Relatives of Annual Averages	598
91. Moving Average Method: Test for Seasonal Pattern of Relatives of Moving Averages	602
92. Link-Relative Method: Test for Seasonal Pattern of Relatives of Preceding Month	605
93. Ratio-to-Trend Method: Test for Seasonal Pattern of Relatives of Trend	610
94. Seasonal Patterns of Cigarette Consumption According to Four Methods	613
95. Relative Cycles of Cigarette Consumption, Monthly, 1927-36	619
96. Successive Steps in the Analysis of a Time Series	637
97. Relative Cycles of Industrial Stock Prices and Commercial Paper Rates, Monthly Data, 1919-37	655
98. Annalist Index of Business Activity	668
99. Babson Chart of Business Conditions	673
100. Buffalo Index of Business Activity	676
101. Eastman Kodak Co.—Planning Chart for Product "S"	693
102. Eastman Kodak Co.—Planning Chart for Product "C"	697
103. Three Scattergram Patterns	706
104. Freehand Regression Line Showing Relation Between Prices and Earnings per share of Common Stock of Twelve Chemical Manufacturers	707
105 A-B-C. Regression Line Fitted by Least Squares Method to Prices and Earnings per share of Common Stock of Twelve Chemical Manufacturers	709-12-13
106. Standard Error of Estimate and Standard Deviation of y , Prices in Relation to Earnings Per Share of Common Stock of Twelve Chemical Manufacturers	718
107. Rates Charged by Banks for Customer Loans in Eight Northern and Eastern Cities Exclusive of New York City, and the Yield of Aaa Bonds, with Straight Line Trend for Each Series, 1919-37	737

FIGURE	PAGE
108. Curves of Three Binomial Expansions Compared with Normal Curve	757
109. Normal Curve Plotted by Calculating Values of Ordinates . . .	764
110. Binomial and Normal Distributions Fitted to Frequency Distribution of Monthly Cost of Electric Current	768
111A. Binomial Frequency Distributions, $N(q+p)^{10}$, for Various Values of q and p when $N = 100$	772
111B. Binomial Frequency Distributions, $N(q+p)^n$, for Various Values of n when $q = .9$, $p = .1$, and $N = 100$	773
112. Diagram for Finding Values of P Associated with Computed Values of χ^2 and $N - m$	778
113. Diagram for Finding Values of $2P$ Associated with Computed Values of t and $N - m$	807
114. Diagram for Finding the Value of z Associated with $2P = .05$ for a Given $N_1 - m_1$ and $N_2 - m_2$	812
115. Diagram for Finding the Value of z Associated with $2P = .01$ for a Given $N_1 - m_1$ and $N_2 - m_2$	813

CHAPTER I

STATISTICS IN BUSINESS

THE STATISTICAL APPROACH

WHEN MASSES of numerical information are to be analyzed some means of summarization must be found which will focus attention upon their major characteristics. Statistical methods have been developed to meet this need; hence in a broad sense the statistical approach is essentially a process of classification, subclassification, and cross-classification designed to give meaning to a mass of information by separating it into comparable parts. Statistical methods therefore are useful in any field of knowledge in which the recording of events produces masses of numerical information. The more important fields are psychology, sociology, education, medicine, biology, public affairs, economics, and business.

Statistical Data Distinguished from Abstract Numbers

Not all numbers are statistics. A table of logarithms is not a statistical table, but simply a compilation of abstract numbers obeying a fixed law. On the other hand statistical data are concrete numbers representing objects or measurements grouped according to stated characteristics. For example in Table 1 sales of low-priced automobiles are classified by make of car and by year of production. This double classification permits comparison of sales of the three makes in any year and the

TABLE 1
SALES OF PASSENGER AUTOMOBILES DURING THE MODEL YEAR 1937-39:
THREE MAKES IN THE LOW-PRICED FIELD*

MAKE OF AUTOMOBILE	MODEL YEAR		
	1937	1938	1939
Chevrolet	804,350	465,403	577,986
Ford	807,258	345,244	456,792
Plymouth	500,503	268,436	387,452
Total	2,112,111	1,079,083	1,422,230

* Compiled from *The Annalist*, Vol. 49, No. 1258, p. 350; Vol. 51, No. 1310, p. 304; Vol. 53, No. 1362, p. 305; Vol. 55, No. 1418, p. 433.

comparison of sales of each with the total. The changes in individual and total sales from year to year can also be read from the table.

Statistics deals with numbers not merely as such, but as the expression of a quantitative or qualitative relationship of the concepts with which they are associated. Statistical work is for the most part a matter of expressing these relationships in the best form, and of finding new relationships. Thus the comparisons observed in Table 1 might be facilitated by the computation of per cent distributions and index numbers. The development of such techniques of analysis forms a major part of the content of subsequent chapters.

Statistics in the Field of Business

While the statistical procedures useful in the several fields of knowledge are in the main identical, those procedures must be adapted to the particular types of information found in each field. The use of ratios is a basic method of analysis common in all types of statistical work but the emphasis on different kinds of comparison varies markedly from one field to another. In vital statistics the study of death rates leads to the development of crude rates, specific rates, standardized rates, and corrected rates. On the other hand business data require per cent relations, per cents of change, per capita ratios, per cent distributions, and index numbers. Whether used in vital statistics or business statistics the word "ratio" implies a relation between two items one of which is the numerator and the other the denominator. But the examples cited show the variation in usage and suggest the extent to which subject matter determines what type of ratio comparisons will be emphasized.

The relation of emphasis to subject matter can be illustrated further by considering time-series analysis. The business statistician spends a major fraction of his time in separating time series into their several components primarily to segregate the cyclical fluctuations. In such fields as medicine, psychology, education, and biometry the techniques of time-series analysis are relatively unimportant and when used seldom have as a goal the study of cyclical fluctuations. In this illustration, as in the preceding one, subject matter and purpose determine to a large extent the form of use and the importance of a particular method of analysis.

These examples should be sufficient to indicate that a development of statistical methods applicable to a particular field of knowledge becomes more specific than a general presentation, and therefore is better suited to the needs of those interested in that field. Consequently this book is devoted in the main to a presentation of statistical methods and operating techniques suitable for the analysis of masses of numerical information arising in the field of business.

The word "business" is taken to include the aggregate of activities involved in transforming raw materials into finished consumable products and transferring goods at all stages of the process. The usual divisions of the field are production, marketing, financial operations, and transportation and communication. Such activities as legal advising, accounting, and technical research including statistical work have not been listed as divisions of business because they are adjunct to all phases of business operation. Statistics in particular may assist in the solution of problems arising in any part of the business field but has its greatest usefulness when large masses of numerical information are to be analyzed.

The following illustrations of the uses of statistical methods in the four main divisions of business enterprise should give sufficient evidence of the pervasiveness of the statistician's work.

Production

- Preparation of production schedules

- Determination of distribution of sizes in manufacture of hats, shoes, suits, dresses, etc.

- Analysis of time and motion studies

- Cost analyses

Marketing

- Determination of sales areas and sales quotas

- Study of effectiveness of advertising

- Relation of size of orders to net profits

- Relation of markdowns of goods to buying policies

Financial operations

- Ratio analysis of financial statements by banks to determine the credit risk of prospective borrowers

- Determination of the average discount rate of customer loans of a bank

Transportation and communication

- Ratio analysis of railroad traffic to determine operating efficiency, operating density, etc.

- Study of relative costs of moving freight by truck and by rail

- Study of telephone and telegraph message density

THE WORK OF THE STATISTICIAN

Types of Statistical Work

The type of problem with which a statistician deals is determined by his location in the economic structure. If he is employed by a business concern his work consists mainly in the analysis of problems arising within the concern and his data are usually the records of the concern's operations. The extent and character of statistical work carried on within any individual concern depend upon the type of business and the funds available. There are many firms that do not maintain separate statistical departments, but which conduct statistical analysis as an adjunct to the main function of one or several departments. Considerable information concerning the variation in statistical practice in different concerns can be obtained from a survey made by the National Industrial Conference Board.

The Conference Board survey, which was begun in June, 1939, reveals that no uniform practice is followed in the organization of research. Only 30% of the companies maintain a separate centralized department for such work, or place the responsibility in the hands of a single statistician or economist. Ten per cent of the concerns assigned such research to a single executive. The majority, about 60%, divided the task among several executive offices and departments.

Somewhat greater centralization of research is found in financial and public service companies than in manufacturing concerns.

The greatest volume of work appears to be done in the accounting and controllers' departments, and the second heaviest volume falls on the executive offices. Next in importance come the sales division, the production department and, in fifth rank, the centralized statistical or economic research department.

Most companies compile data for internal use on sales and orders, production, employment and purchases. Less than one-fourth of the companies reporting, however, attempt to compile data on inventories in the hands of distributors of their products. Most companies also attempt to forecast future trends in sales, production, costs, inventory requirements, prices and profits. More than half of the organizations reported that they attempt to forecast sales by geographic regions.

Forty-two per cent of the companies carrying on research compile periodic reports on the outlook for the particular industry in which they are engaged, and nearly as many compile data on the prospects for business in general. About 15% also prepare studies on general business conditions as they affect purchasers of their finished products and suppliers of their raw materials.

Other special studies carried on to a considerable extent by private industry, listed in the order of their importance, are the economic effects of taxation,

analysis of departmental operations, personnel practices, plant layout, effects of legislation, and feasibility of plant expansion.¹

While this study shows the ramifications of statistical work in business it does not emphasize the variety of problems encountered by an individual statistician. He may be asked to make cost analyses and forecasts of production for the manufacturing department, analyses of time and motion studies for the plant scheduling department, studies of employment, payroll, and wages for the personnel department, sales quotas for the sales department, estimates of plant burden for the maintenance department, studies of bad debt losses for the credit department, or an investigation of the relation between selling prices, sales volume, and turnover of inventory for the president.

The variety of subjects included in this list demonstrates the scope of the work of the statistician employed by an individual concern. He must have considerable familiarity with the operations carried on in all departments of the concern and an understanding of the economic principles involved in those operations, in addition to a working knowledge of statistical methods.

Problems of a different type are dealt with by a statistician engaged in independent research or employed by a trade association, a commercial research agency, a government bureau, or a university research department. Much of the information used in this kind of statistical analysis is gathered from the records of individual concerns or agencies. The data are therefore of essentially the same nature as those used by each individual concern in analyzing its own problems, although the purpose of the analysis is different.

Table 2 is an illustration of a study that makes use of the records of a number of concerns. The Bureau of Business Research of the Harvard Graduate School of Business Administration maintains a regular reporting service through which it receives annual reports of operations from a large number of department stores in all sections of the country. This table gives a summary of the turnover rates computed from the reports of 430 stores. The stores are divided into ten groups according to size as measured by annual sales. The purpose of making this classification is to group together stores operating under conditions that are as nearly similar as possible.

¹ *Commercial and Financial Chronicle*, Vol. 150, No. 3894 (February 10, 1940), p. 904 (New York: William B. Dana Co.), reproduced from a National Industrial Conference Board Report.

TABLE 2
TURNOVER OF GOODS IN DEPARTMENT STORES OF DIFFERENT SIZES IN THE
UNITED STATES, 1938 *

SIZE OF STORE AS MEASURED BY ANNUAL SALES	NUMBER OF STORES REPORTING	AVERAGE TURNOVER OF GOODS
Less than \$150,000	54	2.1
150,000- 300,000	45	2.7
300,000- 500,000	58	3.6
500,000- 750,000	35	3.6
750,000- 1,000,000	28	4.2
1,000,000- 2,000,000	62	4.2
2,000,000- 4,000,000	58	4.4
4,000,000-10,000,000	57	4.7
10,000,000-20,000,000	20	4.7
20,000,000 or more.....	13	5.4

* Malcolm P. McNair, "Operating Results of Department and Specialty Stores in 1938," *Bureau of Business Research Bulletin Number 109* (May, 1939), Boston: Graduate School of Business Administration, Harvard University.

The turnover is computed by dividing the annual sales by the average inventory. The increase in the turnover, as size of store increases, indicates that the smaller stores maintain larger inventories in relation to sales than the larger stores. This skeleton fact gives rise to many questions related to the analysis of department-store operations. For example, one might theorize as follows: the smaller stores must keep in stock a line of goods practically as inclusive as that maintained by larger stores; however, in smaller stores, demand for many types of goods is only occasional, whereas those same goods are in constant demand in larger stores; consequently the maintenance of this slow-moving stock reduces the turnover of the smaller stores. The testing of this hypothesis would be a task for the statistical staff that has access to the reports of the individual concerns.

Further examples of the type of research undertaken by statisticians working with the records of individual concerns are: the relation of advertising costs to sales; the seasonal variation in automobile sales in different parts of the country; the relation of bank loans to size of banks and population of cities in which the banks are located; and the rates of interest charged for installment credit according to type of goods purchased. In other cases the data used in research work do not come from business concerns but from markets, individuals, or the results of prior statistical work. Some illustrations are the construction of an index of the general price level, a study of the preferences of consumers for competing products, and an analysis of the relation

of the alternations of prosperity and depression to the sales of consumers' goods and producers' goods.

The studies mentioned in preceding paragraphs give some indication of the difference in type of problem encountered by statisticians working for individual concerns and by those who engage in business research in some other capacity. Specific techniques that are important in one type of work may be less so in the other, but a common body of method is used in either case. Since the primary purpose of this book is to present a systematic development of statistical method, no occasion will arise for keeping the two types of statistical problems separate. Examples will be drawn from either to illustrate the discussion of methods employed in both.

Statistical Background of Business Activity

The extent to which business operations depend upon a background provided by statisticians is not generally realized. This statement is equally applicable to every technical specialist who is a part of the business structure, but the obscurity of the statistician's contribution is particularly striking because of the wide ramifications of his work.

The nature of the work of the statistician can be explained with the aid of some published statements concerning business affairs.

Example 1.—"Wages [in 1938]" *Swift and Company Year Book* (1938), p. 26:

Since 1923 the average hourly wage rate for Swift & Company's Chicago plant workers has increased by 52 per cent, while the number of hours in the basic working week has been reduced from 48 to 40. Actual weekly earnings per worker are about 37 per cent greater than in 1923. Taking into account the changes in living costs, these weekly earnings provide Swift & Company plant workers with approximately 57 per cent higher "real" wages than they received in 1923.

The statistical department of Swift and Company presumably maintains employment records containing average hourly wages, the number of hours worked per week, and the average weekly earnings per worker in 1923 and in 1938. The computation of the per cents of increase is, of course, routine work. Indexes of the cost of living are published by the United States Bureau of Labor Statistics and the National Industrial Conference Board. A comparison of weekly wages of Swift and Company employees with a cost of living index gives the increase in real wages.

Example 2.—"Your Food Supplies and Costs," *Consumers' Guide*, Vol. V, No. 10 (October 24, 1938), p. 16:

EGGS. Supplies are expected to continue smaller than a year earlier during the remainder of 1938, but in 1939 supplies probably will be bigger than in the current year. Relatively small stocks of storage eggs, coupled with smaller fresh egg production than in 1937 have been the major factors behind the larger than usual price upswing this year. Storage stocks are an important source of supply during the last quarter of the year when fresh egg production reaches its lowest level. Current storage stocks are almost a third under a year ago.

Continuation of the present rate of increase in prices would result in peak egg prices in November considerably above their 1937 level and might result in the highest prices since 1930. There is some possibility, however, that fresh egg production may comprise a larger than usual proportion of total supplies during the last two months of the year because of the large hatchings this spring. This condition would offset part of the price boosting effect of small storage stocks. Retail egg prices went up 5 cents a dozen from August to September and were a cent a dozen higher than last September.

The statistical background for this analysis of egg prices has been supplied by the United States Department of Agriculture. Local offices of the department in all parts of the country send regular reports to Washington concerning conditions in their areas. The analysis of these reports by the statistical division provides information concerning the supply of eggs for the latter part of 1938 and early 1939, the size of cold storage holdings, the prices of eggs during the year, and the prospective supply of egg-laying pullets. Previous studies of the department afford a basis for the statement that "storage stocks are an important source of supply during the last quarter of the year when fresh egg production reaches its lowest level." Comparisons of current reports with department records show absolute and relative price changes from earlier months as well as earlier years.

Example 3.—*Buffalo Evening News* (November 22, 1938), p. 29:

YULE SALES MAY EQUAL \$1,200,000,000 OF 1937

NEW YORK, Nov. 22 (AP).—A busy Christmas shopping season was foreseen today by the National Retail Dry Goods Association.

An analysis by its accounting experts, the association reported, indicated dollar sales in department and apparel specialty stores of the nation in the four weeks preceding Christmas may approximate \$1,200,000,000, about the same as in the comparable 1937 period.

Actually, the number of items traded across store counters, it was pointed out, may exceed last year's Christmas trade because department store prices this year are about 7 per cent lower on the average.

As stated in the article the accountants have estimated that sales during the Christmas season of 1938 will be very satisfactory, but the main point is the work of the statistician which is back of the innocent-looking statement that prices of department store goods are about 7 per cent lower than they were in 1937. This conclusion is probably based on the "Index of Prices of Department Store Goods" prepared monthly by A. W. Zelomek and published in *Fairchild Publications*. This index is based on prices of 105 nonstyle items collected monthly from 53 retail-trade organizations.

Example 4.—"The Trend of Business," *Dun's Review*, Vol. 46, No. 2121 (May, 1938), pp. 30-31:

On the charts, the present state of business activity bears some resemblance to that of 1934. A few of the more important measures—national income, department store sales, wholesale prices, construction contracts—are still above early 1936 levels, considerably higher than in 1934. On the other hand, industrial production is down to the 1934 average; primary distribution, measured by railroad carloadings, is the lowest since November, 1934; the *Annalist* index of business activity for March, the lowest since November, 1934; the *Times* average of 50 stock prices for the first three weeks of April, the lowest since September, 1934.

The first sentence indicates that charts have been prepared by statisticians showing the course followed by various indicators of business conditions in recent years. The computation of the national income requires the continuous attention of a corps of statisticians in the United States Department of Commerce. Department-store sales are reported by over 400 individual stores to the Federal Reserve Banks of the districts in which the stores are located. Indexes of sales are prepared for each district as well as for the United States as a whole. Wholesale price indexes are prepared by a number of statistical agencies, but the most widely used index is that of the United States Bureau of Labor Statistics computed by an elaborate technique and based on prices of over 800 commodities. Data on construction contracts are collected by the F. W. Dodge Corporation through local offices and correspondents in 37 states east of the Rocky Mountains. *An Index of Industrial Production* is published by the Board of Governors of the Federal Reserve System. The research staff prepares the

index by the application of elaborate statistical techniques to data compiled from trade journals, reports of trade associations and government bureaus. Railroad car loadings are collected from individual railroad companies and prepared for publication by the Car Service Division of the Association of American Railroads. The *Annalist* index of business activity is a cyclical index corrected for trend and seasonal variation by an involved statistical process. The *New York Times* average of 50 stocks is a product of the newspaper's research staff.

Articles similar to these are presented every day to the reading public and they exert a widespread influence over the conduct of business affairs. These four examples give some indication of the variety of the activities of business statisticians and of the multiplicity of methods and techniques they employ. The orderly development of basic methods and techniques and their relation to various business activities become the subject matter of a textbook in statistics.

PROBLEMS

- 1. What distinguishes statistical data from abstract numbers?
- 2. Apply this distinction to the following; give reasons for your answer in each case:

A

NUMBERS	SQUARES	SQUARE ROOTS	RECIPROCALs
51.....	2601	7.1414	.019608
52.....	2704	7.2111	.019231
53.....	2809	7.2801	.018868
54.....	2916	7.3485	.018519
55.....	3025	7.4162	.018182

B

REPORT OF OVERTIME WORKED BY LOCAL BRANCHES OF A LABOR UNION	
AMOUNT OF OVERTIME	No. OF LOCALS
None	2
Occasionally	5
Never more than 6 hours per week.....	1
When necessary	3
Five hours regularly.....	2
Total	13

C

ADDITIONS TO TERRITORY OF CONTINENTAL UNITED STATES AFTER 1783	
TERRITORY	DATE OF ADDITION
Northwest Territory	1787
Louisiana Purchase	1803
Florida	1819
Texas	1845
Oregon	1846
Mexican Cession	1848
Gadsden Purchase	1853

D

HOURLY UNSKILLED HIRING WAGE RATE OF A GROUP OF MANUFACTURING CONCERNS
IN 1936

CONCERN	HOURLY WAGE RATE (in Cents)
A	32
B	36
C	30
D	40
E	35

3. What are the differences between the study of general statistics and business statistics?
4. Why is statistics not listed as a division of business activity?
5. List the differences in function of the statistician employed by a private concern and one employed by some other type of organization.
6. In the preparation of which of the following reports would the statistician initially compiling the data be employed by an industrial concern?
 - a) Monthly production of automobiles and trucks by General Motors.
 - b) Weekly freight car loadings of coal in the United States.
 - c) Daily bank clearings of the Buffalo Clearing House Association.
 - d) The monthly production of crude oil in the United States.
 - e) The net profits of the Erie Railroad for the first six months of 1930.
 - f) The daily messages carried by the New York Telephone Company.
 - g) The number of airplanes arriving and departing at the Buffalo airport.
 - h) Number of bank employees:

NAME OF BANK	EXECUTIVES	CLERKS
Export	23	180
First National,	14	200
etc.	...	...
Total	212	1325

7. Describe the statistical material found on the financial pages of an urban newspaper. Be sure to give exact reference to the issue and edition of the paper.
8. Select from a current publication an article similar to the examples in the text. State what work has been done by statisticians in the preparation of the article. Give exact reference.

REFERENCES

BROWN, THEODORE H., "Problems Met by Companies That Instruct Their Employees in Statistical Methodology," *Journal of the American Statistical Association*, Vol. XXVII, No. 181A (March, 1933, Supplement), pp. 10-14.

A suggested program of statistical work that could be carried on within a business concern.

BURGESS, ROBERT W., "The Whole Duty of the Statistical Forecaster," *Journal of the American Statistical Association*, Vol. XXVII, No. 181A (March, 1933, Supplement), pp. 636-42.

The first part of this article includes some excellent examples of the types of analysis made by statisticians.

FALKNER, ROLAND P., "The Scope of Business Statistics," *Quarterly Publications of the American Statistical Association*, Vol. XVI, No. 122 (June, 1918), pp. 24-29.

An earlier attempt to define business statistics that is still pertinent.

HATHAWAY, WILLIAM A., "Internal and External Statistical Needs of American Business," *Quarterly Publications of the American Statistical Association*, Vol. XVI, No. 122 (June, 1918), pp. 1-15.

Contains background essential to an understanding of modern statistical development.

PARMALEE, JULIUS H., "The Utilization of Statistics in Business," *Quarterly Publications of the American Statistical Association*, Vol. XV, No. 117 (June, 1917), pp. 565-76.

An early statement of the types of statistical analysis available to business men.

YOUNG, BENJAMIN A., *Statistics as Applied in Business*. New York: The Ronald Press Company, 1925.

Chapters I to VIII contain a very complete statement of the character of internal and external statistical work and a detailed presentation of types of statistical problems.

CHAPTER II

THE USE OF NUMBERS

INTRODUCTION

IN BUSINESS practice there is an increasing trend toward the expression of ideas in numerical form. The manager of a store no longer reports that business is improving, but that sales last month were 16 per cent better than in the corresponding month of last year. The banker no longer relies solely upon his personal judgment in granting loans, but uses a set of ratios, derived from the financial statement of a prospective borrower, to aid him in determining the concern's credit standing. A similar tendency toward more precise methods can be found in all parts of the business structure. In no small degree this tendency accounts for the increasing demand for a knowledge of statistical methods.

On the other hand teachers in various parts of the country have remarked that young men and women of college age show a decline in ability to carry out numerical operations and particularly an increasing inability to think in numerical terms. It is not the function of a textbook in statistics to reverse this tendency. The fault is too fundamental for that. This condition does explain, however, why it is desirable to pause for a brief statement concerning methods of computation before proceeding with the development of statistical techniques.

The necessary computation which accompanies statistical work consumes a vast amount of time. The greater part of such computation is purely repetitive in character; consequently methods of shortening the time spent in doing it will allow more of the student's time to be spent in studying statistics and less in practicing arithmetic. Hence the following pages are devoted to a review of arithmetic operations.

THE FUNDAMENTAL OPERATIONS

Addition

Computations should be performed rapidly. The advantage in speed lies not merely in the time saved, but mainly in the confidence

gained for those who waver and in the attention preserved for those whose minds might wander. There is some advantage here in illustrating the wrong method. Suppose that the following columns are to be added:

$$\begin{array}{r}
 2641 \\
 362 \\
 570 \\
 1369 \\
 8147 \\
 4216 \\
 \hline
 2164
 \end{array}$$

All too commonly in the author's experience the student's mind goes through the following steps: adding up from the bottom, 4 and 6 are 10, 10 and 7 are 17, 17 and 9 are (then 9 are told off on the fingers) are 26, 26 and 2 are (I think I'll go to lunch after this class). Let me see, I was adding something. Oh yes, 4 and 6 are 10, etc.

To eliminate this wandering the addition should be done at the maximum speed possible, naming only the successive sums. So, first column, 10, 17, 26, 29; second column, 9, 19, 26, 36; third column, 6, 10, 15, 24; fourth column, 8, 16, 19. There is everything to be gained and nothing to be lost by performing addition at a rate of speed which will leave no time to worry about an impending lunch hour. There are students who can be busily engaged for as much as 2 minutes in adding these columns of figures, whereas 15 seconds is the maximum time which should be spent.

Those who have difficulty with the amount to be carried from one column to the next may prefer to write the total of each column separately as indicated below.

$$\begin{array}{r}
 \hline 29 \\
 34 \\
 21 \\
 17 \\
 \hline 19469
 \end{array}$$

This method is advantageous if one is likely to be interrupted.

Subtraction

The results of subtraction should always be checked by adding the subtrahend and the remainder.

$$\begin{array}{r}
 38642 \text{ minuend} \\
 (-) 12966 \text{ subtrahend} \\
 \hline
 25676 \text{ remainder} \\
 (+) 12966 \text{ subtrahend} \\
 \hline
 38642 \text{ check}
 \end{array}$$

Subtraction of numbers can be performed mentally by using a method of excess and deficit in relation to a round number such as 100. In subtracting 69 from 118, the minuend is $100 + 18$, the subtrahend is $100 - 31$; therefore the difference is $18 + 31 = 49$. Other similar examples are:

$$\begin{array}{r}
 263 = 200 + 63 \\
 - 186 = 200 - 14 \\
 \hline
 77 \text{ remainder}
 \end{array}
 \qquad
 \begin{array}{r}
 547 = 500 + 47 \\
 - 490 = 500 - 10 \\
 \hline
 57 \text{ remainder}
 \end{array}$$

$$\begin{array}{r}
 1245 = 1000 + 245 \\
 - 893 = 1000 - 107 \\
 \hline
 352 \text{ remainder}
 \end{array}
 \qquad
 \begin{array}{r}
 2317 = 2200 + 117 \\
 - 2049 = 2200 - 151 \\
 \hline
 268 \text{ remainder}
 \end{array}$$

After this method has been mastered, it will not be necessary to put any of the computation on paper. With sufficient practice the method can be used in fairly complicated operations. For example,

$$\begin{array}{r}
 12286 = 12000 + 286 \\
 - 1143 = 1000 + 143 \\
 \hline
 11000 + 143 = 11143
 \end{array}
 \left. \vphantom{\begin{array}{r} 12286 \\ - 1143 \end{array}} \right\} \text{subtracting}$$

$$\begin{array}{r}
 4260 = 4000 + 260 \\
 - 2749 = 3000 - 251 \\
 \hline
 1000 + 511 = 1511
 \end{array}$$

Multiplication

The greatest saving of time in multiplication results from the use of short cuts. These are derived from well-known principles of arithmetic and algebra as indicated in the examples of various methods which follow.

*The Use of Reciprocals.*¹—One number multiplied by another is the same as the first number divided by the reciprocal of the second.

1. $763 \times 5 = \frac{763}{.2} = \frac{7630}{2} = 3815$
2. $1582 \times 25 = \frac{158200}{4} = 39550$
3. $220 \times 50 = \frac{22000}{2} = 11000$
4. $17228 \times 125 = \frac{17228000}{8} = 2153500$
5. $15415 \times .16\frac{2}{3} = \frac{15415}{.6} = 2569.17$

¹ The reciprocal of a number is defined as unity divided by the number, i.e., the reciprocal of 5 is $1 \div 5 = .2$. The reciprocal of 40 is $1 \div 40 = .025$. The reciprocal of .25 is $1 \div .25 = 4$.

Multiplier Near Ten or a Power of Ten.—

6. $27 \times 99 = 27(100-1) = 27(100) - 27(1) = 2700 - 27 = 2673$
7. $366 \times 1001 = 366000 + 366 = 366366$
8. $2746 \times 11 = 27460 + 2746 = 30206$

Squaring Numbers Ending in Five.²—

9. $25^2 = (2 \times 3)100 + 25 = 625$, i.e., $2(2+1)$ and annex 25
10. $75^2 = (7 \times 8)100 + 25 = 5625$, i.e., $7(7+1)$ and annex 25
11. $105^2 = (10 \times 11)100 + 25 = 11025$ i.e., $10(10+1)$ and annex 25
12. $405^2 = (40 \times 41)100 + 25 = 164025$, i.e., $40(40+1)$ and annex 25

Last Digits Totaling Ten.³—

13. $67 \times 73 = (70-3)(70+3) = 4900 - 9 = 4891$
14. $95 \times 105 = (100-5)(100+5) = 10000 - 25 = 9975$
15. $89 \times 111 = (100-11)(100+11) = 10000 - 121 = 9879$
16. $4.1 \times 2.9 = (3.5+.6)(3.5-.6) = 12.25 - .36 = 11.89$
17. $640 \times 5.6 = 100(6.4 \times 5.6) = 100(6+.4)(6-.4) = 100(36-.16) = 100 \times 35.84 = 3584$

A Method of Squaring Any Number.⁴—

18. $72^2 = (72-2)(72+2) + 2^2 = (70 \times 74) + 4 = 5180 + 4 = 5184$
19. $153^2 = (153-3)(153+3) + 3^2 = (150 \times 156) + 9 = (100 \times 156) + (50 \times 156) + 9 = 15600 + 7800 + 9 = 23409$

Division .

There are a few worthwhile short cuts in division, based mainly on the use of reciprocals.⁵ Commonly used among these are,

20. $5725 \div 25 = 5725 \times .04 = 57.25 \times 4 = 229$
21. $280400 \div 50 = 2804 \times 2 = 5608$
22. $12750 \div 500 = 12.75 \times 2 = 25.5$
23. $245925 \div 125 = 245.925 \times 8 = 1967.4$

It is to be expected that the reader who employs the few short cuts listed here will develop many more to aid him as he progresses in the

² To square any number ending in 5 multiply the part of the number to the left of 5 by one more than itself and annex 25 to the product.

³ Examples 13 to 17 make use of the algebraic identity $(a+b)(a-b) = a^2 - b^2$.

⁴ The same algebraic identity is used as in the preceding examples but the form is changed to $a^2 = (a+b)(a-b) + b^2$.

⁵ The use of reciprocals changes division to multiplication whereas the use of reciprocals on a preceding page changed multiplication to division.

use of numbers. There is no standard set which can be recommended for the use of everyone. Each person should employ those which result in time saving for him and come to mind naturally. Just as every person has an individual style in writing so every person will develop an individual set of short cuts in computation.

The Order of Performing the Fundamental Operations

When the operations of addition and subtraction are employed in a problem, the order of performing them makes no difference in the result. Thus:

$$\begin{array}{rcl} & 50 + 275 - 36 + 5 - 210 = 84 \\ \text{or} & 275 - 36 + 50 - 210 + 5 = 84 \\ \text{or} & -210 + 50 - 36 + 275 + 5 = 84 \end{array}$$

The introduction of parentheses into such a series indicates that the operations within the parentheses must be performed first. There will be no difference in the result when the sign preceding a parenthesis is plus, but when it is minus, it has the effect of reversing the sign of every figure inclosed. Thus:

$$\begin{array}{l} 69 - 63 + 58 - 10 = 54 \\ 69 - (63 + 58) - 10 = 69 - 121 - 10 = -62 \\ 69 - (63 + 58 - 10) = 69 - 111 = -42 \end{array}$$

but

$$69 - 63 + (58 - 10) = 69 - 63 + 48 = 54$$

When the operations of multiplication and division are employed in a problem, the order of performing the division does alter the result; hence in order to avoid ambiguity, it is necessary to inclose in parentheses the figures that are intended to be used together as numerator and denominator. Thus:

$$(250 \div 10) \times 2 = 25 \times 2 = 50$$

but

$$250 \div (10 \times 2) = 250 \div 20 = 12.5$$

If several signs of grouping are used in the same problem, the rule is "Work from the inside out." Thus:

$$\begin{aligned} -5 \{ [26 (36 \div 9)] \div 52 \} &= -5 \{ [26 \times 4] \div 52 \} = -5 [104 \div 52] \\ &= -5 \times 2 = -10 \end{aligned}$$

When multiplication or division or both appear in a problem along with addition or subtraction or both, with no parentheses, the opera-

tions of multiplication and division must be performed first. If parentheses are introduced, the rules already quoted will apply. Thus:

$$\begin{aligned}
 550 + 10 \times 7 - 60 \div 5 &= 550 + 70 - 12 = 608 \\
 (550 + 10) \times (7 - 60) \div 5 &= 560 \times (-53) \div 5 = -29680 \div 5 \\
 &= -5936 \\
 (550 + 10) \times [7 - (60 \div 5)] &= 560 \times [7 - 12] = 560 \times (-5) \\
 &= -2800
 \end{aligned}$$

FRACTIONS

Sometimes computations are carried on in common fractions and sometimes in decimals. It is desirable therefore to know how to perform operations with both and how to convert one into its equivalent in the other.

Common Fractions

Addition and Subtraction.—To add $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{5}$, the common denominator must first be found. The common denominator is the smallest number divisible by the individual denominators, in this case 2, 3, and 5. By inspection this is 30; there is no smaller number divisible by 2, 3, and 5. The three fractions with the common denominator 30 are $\frac{1}{2} \times \frac{15}{15} + \frac{1}{3} \times \frac{10}{10} + \frac{1}{5} \times \frac{6}{6} = \frac{3}{30} + \frac{2}{30} + \frac{1}{30} = \frac{6}{30}$ or $1\frac{1}{5}$. Suppose that the four fractions $3\frac{3}{8}$, $5\frac{4}{15}$, $2\frac{3}{20}$, and $8\frac{2}{9}$ were to be added. When, as in this case, the common denominator is not evident by inspection, the general method of finding it is to reduce all of these individual denominators to their prime factors and take the product of the prime factors appearing in the reduction. The form for finding the common denominator of the given fractions is:

Divisors	Denominators			
2	36	15	20	9
2	18	15	10	9
3	9	15	5	9
3	3	5	5	3
5	1	5	5	1
	1	1	1	1

The process consists in dividing by 2 as long as any of the denominators are divisible by 2. Then do the same with 3 and so on using only

prime⁶ numbers as divisors until unity is reached in each column. When any denominator is not divisible in any row, it is simply carried along until a divisor is used by which that denominator is divisible. The common denominator will be the product of the prime divisors on the left, i.e., $2 \times 2 \times 3 \times 3 \times 5 = 180$. The four fractions expressed in terms of the common denominator are, $3\frac{65}{180} + 5\frac{48}{180} + 2\frac{27}{180} + 8\frac{40}{180} = 18\frac{180}{180} = 19$.

Subtraction of fractions is performed by the same process of reducing to a common denominator. For example, in subtracting $6\frac{5}{8}$ from $13\frac{9}{5}$, both fractions should be changed to the common denominator 90. Then $13\frac{9}{5} - 6\frac{5}{8} = 12\frac{108}{90} - 6\frac{56}{90} = 6\frac{52}{90}$.

Multiplication.—The type of multiplication problem most common in statistical work involving common fractions is similar to that met by the bookkeeper in extending bills or inventories, e.g., $4\frac{5}{8}$ dozen shirts at $\$16\frac{1}{2}$ per dozen. Reduce each number to an improper fraction: $4\frac{5}{8} = \frac{29}{8}$ and $16\frac{1}{2} = \frac{33}{2}$; then the total value of the shirts would be $\frac{29}{8} \times \frac{33}{2} = \frac{29}{2 \times 8} \times \frac{33 \times 11}{2} = \frac{29 \times 11}{4} = \frac{319}{4} = \$79\frac{3}{4}$.

Division.—The rule for finding the value of the quotient of two fractions is: invert the fraction in the denominator and multiply. Some examples are:

$$\begin{aligned}\frac{3}{8} \div \frac{1}{5} &= \frac{3}{8} \times \frac{5}{1} = 3 \\ 4\frac{5}{8} \div 1\frac{1}{4} &= \frac{33}{8} \div \frac{5}{4} = \frac{33}{8} \times \frac{4}{5} = \frac{33}{20} = 3\frac{9}{20} \\ \frac{2\frac{5}{11} \times 7\frac{3}{5}}{2\frac{4}{5}} &= \frac{\frac{27}{11} \times \frac{38}{5}}{\frac{14}{5}} = \frac{9 \times 38}{11} \times \frac{5 \times 16}{8} \times \frac{1 \times 1}{16} = 9\end{aligned}$$

Decimal Fractions

The preceding paragraphs have dealt with problems containing common fractions. It is equally necessary to be conversant with methods of dealing with decimal fractions. In fact decimals are more frequently employed in statistical work than common fractions. The use of calculating machines requires the expression of fractional amounts decimally. It is necessary therefore that statisticians acquire facility in handling both common and decimal fractions and be able to convert one to the other automatically. To convert a common fraction to a decimal the numerator is divided by the denominator, i.e.,

$$\begin{aligned}\frac{1}{5} &= 1.0 \div 5 = .2 \\ \frac{3}{8} &= 3.000 \div 8 = .375 \\ \frac{5}{140} &= 5.000 \div 140 = .0357 \dots\end{aligned}$$

⁶ A prime number is one that is divisible only by unity and the number itself.

Decimal Fractions and Per Cents.—When any number is expressed as a decimal or a per cent, it simply means that the numerator of a common fraction is written, the denominator being understood without writing it. Decimals mean a certain part of one unit, while per cents mean a certain part of 100 units. Thus, .5 means five-tenths of one unit, or one-half, while 50 per cent means 50 of 100 units. Obviously $\frac{1}{2}$ of every one and 50 of every hundred are merely two ways of expressing the same relation; hence we say that .5 is equivalent to 50 per cent. The rule is: *To express a decimal as a per cent move the decimal point two places to the right.* The reverse rule is: *To express a per cent as a decimal move the decimal point two places to the left.*

It is important for the statistician to be able to change from common fractions or decimals to per cents and the reverse with accuracy and without consuming much time in the process. Table 3 gives a list of equivalents that can be referred to until their use becomes automatic.

TABLE 3
LIST OF COMMON FRACTIONS WITH THEIR DECIMAL AND PER CENT EQUIVALENTS

COMMON FRACTION	EQUIVALENT DECIMAL	EQUIVALENT PER CENT	COMMON FRACTION	EQUIVALENT DECIMAL	EQUIVALENT PER CENT
$\frac{1}{1000}$..	.001	.1	$\frac{5}{8}$..	.625	62.5
$\frac{1}{500}$..	.002	.2	$\frac{7}{8}$..	.875	87.5
$\frac{1}{400}$	.0025	.25	$\frac{1}{6}$	.166...	16.66...
$\frac{3}{100}$	.00333...	.33...	$\frac{5}{6}$	.833...	83.33...
$\frac{1}{250}$	.004	.4	$\frac{1}{5}$..	.2	20.
$\frac{1}{200}$	.005	.5	$\frac{2}{5}$..	.4	40.
$\frac{1}{125}$	.00625	.625	$\frac{3}{5}$..	.6	60.
$\frac{1}{80}$	.0075	.75	$\frac{4}{5}$..	.8	80.
$\frac{1}{100}$	.01	1.	$\frac{1}{4}$..	.25	25.
$\frac{3}{200}$	.015	1.5	$\frac{3}{4}$..	.75	75.
$\frac{1}{50}$	.02	2.	$\frac{1}{3}$	.33...	33.33...
$\frac{1}{40}$	.025	2.5	$\frac{2}{3}$	.66...	66.66...
$\frac{1}{30}$	.03	3.	$\frac{1}{2}$..	.5	50.
$\frac{3}{100}$	.0333...	3.33...	$\frac{1}{2}$..	1.5	150.
$\frac{1}{25}$	.04	4.	$\frac{1}{3}$..	1.33...	133.33...
$\frac{1}{20}$	.05	5.	$\frac{5}{4}$	1.25	125.
$\frac{1}{125}$	.0625	6.25	$1\frac{3}{4}$..	1.75	175.
$\frac{1}{15}$	.066...	6.66...	$2\frac{1}{2}$	2.2	220.
$\frac{1}{12}$	.0833...	8.33...	$3\frac{3}{4}$	3.625	362.5
$\frac{1}{9}$	.1166...	11.66...	$4\frac{1}{2}$	4.875	487.5
$\frac{7}{12}$	.5833...	58.33...	$5\frac{1}{2}$	5.833...	583.33...
$1\frac{1}{2}$	.9166...	91.66...	$8\frac{1}{10}$..	8.1	810.
$\frac{1}{8}$	.125	12.5	$10\frac{1}{10}$..	10.125	1012.5
$\frac{3}{8}$	.375	37.5	$12\frac{3}{10}$..	12.35	1235.

Calculation of Per Cents.—The three terms of a per cent calculation are: (1) the base, b , (2) the rate, r , and (3) the percentage, p . *The fundamental relation is $b \times r = p$.* Given any two of these terms the third one can be found from the fundamental relation. There are, therefore, three types of problems which arise. Each of these is illustrated in the examples which follow:

$$b \times r = p$$

Example 1: How much is 5 per cent of 12420?

This means that 5 of every hundred in 124.20 hundreds are to be counted, so $5 \times 124.20 = 621$, hence 5 per cent of 12420 is 621.

The simpler way of doing the same thing is to multiply the given number, 12420 by $.05 = 621$. That is, instead of taking 5 out of every hundred in the original number, take $.05$ out of every one in the original number.

Example 2: How much is 364 per cent of 1250?

$$1250 \times 3.64 = 4550.$$

Example 3: How much is 750 increased by 40 per cent of itself?

$$750 + (750 \times .4) = 750 + 300 = 1050, \text{ or } 750 \times 1.4 = 1050.$$

Example 4: How much is $\frac{3}{8}$ of 4875? $\frac{3}{8}$ per cent of 4875?

$$4875 \times .4 = 1950. \qquad 4875 \times .004 = 19.50.$$

$$p \div r = b$$

Example 5: 450 is 75 per cent of what number?

If 75 per cent of a certain number is 450, then 1 per cent of the number is $\frac{1}{75}$ of 450 or 6. If 1 per cent of a number is 6, then 100 per cent of the number is 100 times 6, or 600. Therefore the number is 600. The work can be shortened as follows: $450 \div .75 = 600$.

Example 6: 375 is $12\frac{1}{2}$ per cent of what number?

$$375 \div .125 = 3000.$$

Another solution would use the $12\frac{1}{2}$ per cent as $\frac{1}{8}$. The problem would then read, 375 is $\frac{1}{8}$ of what number? The number is $375 \times 8 = 3000$.

Example 7: 12500 is $\frac{3}{8}$ of what number?

If 12500 is $\frac{3}{8}$ of the number, then $\frac{3}{8}$ of the number would be $\frac{1}{4}$ of 12500 or 3125. If $\frac{3}{8}$ of the number is 3125, then $\frac{3}{8}$ or 100 per cent

of the number would be 3 times 3125, or 9375. Therefore the number is 9375. The pencil and paper solution would be, $12500 \div 1.3333 = 9375$ plus a remainder. This remainder is, of course, due to the use of an approximate divisor. The advantage of the common fraction solution is obvious.

$$p \div b = r$$

Example 8: What per cent of 24 is 3?

This problem may be worked in two ways,

a) 3 is $\frac{1}{8}$ of 24 or $12\frac{1}{2}$ per cent of 24.

b) $3 \div 24 = .125$ or $12\frac{1}{2}$ per cent.

Example 9: What per cent of 8100 is 17415?

$$17415 \div 8100 = 2.15 \text{ or } 215 \text{ per cent.}$$

Although the wording of a problem may somewhat obscure the case, all per cent calculations can be expressed in one of these three forms. With sufficient practice in dealing with per cent problems no difficulty should be encountered in determining which of the three forms is to be used.

SQUARE ROOTS

The five commonly used methods of determining square roots of numbers are (a) by inspection, (b) by arithmetic, (c) by the use of logarithms, (d) by the use of a table of square roots, and (e) by the use of a slide rule. Only the first and second methods will be discussed at this time. The use of logarithms is explained in Appendix C. A table of square roots has been provided in Appendix D. The use of a slide rule can be learned best from the manual of instructions provided by the manufacturers of slide rules.

By Inspection

The approximate value of the square roots of many numbers can be ascertained by a process of mental interpolation, if one has at command the values of the square roots of a few numbers or makes use of the short-cut method of squaring numbers ending in five. The inspection method can be explained easily by the use of an example. Suppose the square root of 457 were wanted. Twenty squared is 400 and twenty-five squared is 625. The square root of 457 is somewhere between 20 and 25, but it is obviously closer to 20. The difference

between 400 and 625 is 225, 57 is approximately one-fourth of this amount, therefore the square root of 457 is approximately $20 + \frac{1}{4}(5)$ or 21.25. The correct value is 21.38, hence the value by inspection is not even correct to one decimal place.

A variation of the method used in the preceding example will yield better results if a calculating machine is available. Suppose the square root of 12750 were wanted. The square of 105 is 11025 and the square of 110 is 12100 and the square of 115 is 13225. The square root of 12750 is between 110 and 115 but is a little closer to 115. Therefore square 113 on the calculating machine, securing 12769. This is rather close, but the next step, if more accuracy is wanted, is to square 112.9. The result 12746 is still closer and further trials will give 112.92 as the correct root to five figures.

The inspection method yields good results quickly after a little practice. Even though it is not used as a method of finding the square root, it is valuable as a checking device when other methods are employed. This is particularly true when roots are found by logarithms, the slide rule, or a square root table.

By Arithmetic Computation

When no auxiliary devices are available and accurate results are required square roots can be found by the following steps.

Step 1.—Divide the number into groups of two digits each way from the decimal point. The last group on the left will contain only one digit if the number has an odd number of digits to the left of the decimal point.

Step 2.—The largest number whose square does not exceed the value of the digit or pair of digits in the left-hand group of the number is the first figure of the root. This figure is entered above the left hand group.

Step 3.—Subtract the square of the first figure of the root from the left group of the number.

Step 4.—At the right of the remainder of Step 3, annex the figures in the second group of the number. This is the new dividend.

Step 5.—Double the root already found and annex one zero to the right as a trial divisor and divide it into the dividend of Step 4 to obtain the second figure of the root which is entered over the second pair of digits. This new figure will often be too high and must be corrected by trial and error.

Step 6.—The new figure is added to the trial divisor of Step 5 to give the true divisor which is then multiplied by the new figure of the root to give a product which is entered under the dividend of Step 4.

Step 7.—This product is subtracted from the dividend, the next group of digits is annexed on the right of the remainder, and the process of Step 5 and Step 6 is repeated.

Two examples of the use of these seven steps in extracting the square root are shown in Figure 1. The examples are constructed to emphasize the growth of the solution as the successive steps of the process are applied to the examples. The complete solution for future use is given in the last computation at the bottom of the Figure.

FIGURE 1

TWO EXAMPLES OF THE DEVELOPMENT BY SUCCESSIVE STEPS OF THE SOLUTION FOR EXTRACTING SQUARE ROOT

Find the square root of 12750

Steps 1, 2 and 3

$$\begin{array}{r} \sqrt{1 \ 27 \ 50.} \\ 1 \\ \hline 0 \end{array}$$

Steps 1, 2, 3
and 4

$$\begin{array}{r} \sqrt{1 \ 27 \ 50.} \\ 1 \\ \hline 27 \end{array}$$

Steps 1, 2, 3,
4 and 5

$$\begin{array}{r} \sqrt{1 \ 1 \ 27 \ 50.} \\ 1 \\ \hline 20 \ 27 \end{array}$$

Steps 1, 2, 3, 4,
5 and 6

$$\begin{array}{r} \sqrt{1 \ 1 \ 27 \ 50.} \\ 1 \\ \hline 20 \ 27 \\ 1 \\ \hline 21 \ 21 \end{array}$$

Find the square root of 4693.49

$$\begin{array}{r} \sqrt{6 \ 93.49} \\ 46 \\ \hline 36 \\ 10 \end{array}$$

$$\begin{array}{r} \sqrt{6 \ 93.49} \\ 46 \\ \hline 36 \\ 10 \ 93 \end{array}$$

$$\begin{array}{r} \sqrt{6 \ 9 \ 93.49} \\ 46 \\ \hline 36 \\ 120 \ 10 \ 93 \end{array}$$

$$\begin{array}{r} \sqrt{6 \ 9 \ 93.49} \\ 46 \\ \hline 36 \\ 120 \ 10 \ 93 \\ 9 \\ \hline 129 \ 11 \ 61* \end{array}$$

* This product is too large, hence the new root should be 8, as shown below:

$$\begin{array}{r} \sqrt{6 \ 8 \ 93.49} \\ 46 \\ \hline 36 \\ 120 \ 10 \ 93 \\ 8 \\ \hline 128 \ 10 \ 24 \end{array}$$

FIGURE 1—Cont.

TWO EXAMPLES OF THE DEVELOPMENT BY SUCCESSIVE STEPS OF THE SOLUTION FOR EXTRACTING SQUARE ROOT

Steps 1, 2, 3, 4,
5, 6 and 7

$$\begin{array}{r|l}
 1 & 1 \ 1 \ 2 \\
 1 & 27 \ 50. \\
 \hline
 1 & \\
 20 & 27 \\
 1 & \\
 \hline
 21 & 21 \\
 220 & 6 \ 50 \\
 2 & \\
 \hline
 222 & 4 \ 44
 \end{array}$$

$$\begin{array}{r|l}
 6 & 8. \ 5 \\
 46 & 93.49 \\
 \hline
 36 & \\
 120 & 10 \ 93 \\
 8 & \\
 \hline
 128 & 10 \ 24 \\
 1360 & 69.49 \\
 5 & \\
 \hline
 1365 & 68.25
 \end{array}$$

Steps 5, 6 and
7 repeated

$$\begin{array}{r|l}
 1 & 1 \ 1 \ 2. \ 9 \\
 1 & 27 \ 50.00 \\
 \hline
 1 & \\
 20 & 27 \\
 1 & \\
 \hline
 21 & 21 \\
 220 & 6 \ 50 \\
 2 & \\
 \hline
 222 & 4 \ 44 \\
 2240 & 2 \ 06.00 \\
 9 & \\
 \hline
 2249 & 2 \ 02 \ 41
 \end{array}$$

$$\begin{array}{r|l}
 6 & 8. \ 5 \ 0\ddagger \\
 46 & 93.49 \ 00 \\
 \hline
 36 & \\
 120 & 10 \ 93 \\
 8 & \\
 \hline
 128 & 10 \ 24 \\
 1360 & 69.49 \\
 5 & \\
 \hline
 1365 & 68.25 \\
 13700 & 1.24 \ 00
 \end{array}$$

† 13700 will not go into 12400, therefore the next figure of the root is 0 and the next group of figures is brought down.

Steps 5, 6 and
7 repeated

$$\begin{array}{r|l}
 1 & 1 \ 1 \ 2. \ 9 \ 1 \\
 1 & 27 \ 50.00 \ 00 \\
 \hline
 1 & \\
 20 & 27 \\
 1 & \\
 \hline
 21 & 21 \\
 220 & 6 \ 50 \\
 2 & \\
 \hline
 222 & 4 \ 44 \\
 2240 & 2 \ 06.00 \\
 9 & \\
 \hline
 2249 & 2 \ 02 \ 41 \\
 22580 & 3.59 \ 00 \\
 1 & \\
 \hline
 22581 & 2.25 \ 81 \\
 & 1.33 \ 19
 \end{array}$$

$$\begin{array}{r|l}
 6 & 8. \ 5 \ 0 \ 9 \\
 46 & 93.49 \ 00 \ 00 \\
 \hline
 36 & \\
 120 & 10 \ 93 \\
 8 & \\
 \hline
 128 & 10 \ 24 \\
 1360 & 69.49 \\
 5 & \\
 \hline
 1365 & 68.25 \\
 137000 & 1.24 \ 00 \ 00 \\
 9 & \\
 \hline
 137009 & 1.23 \ 30 \ 81 \\
 & 69 \ 19
 \end{array}$$

∴ The square root of 12750 is 112.92. (The last digit of the root is increased to 2 because the remainder is more than half of the last divisor.)

∴ The square root of 4693.49 is 68.509

The example on the right shows how to find the new figure in the root by trial and error in Step 5 and how to deal with a zero in the root. There will be one digit to the left of the decimal point in the root for every pair of figures to the left of the decimal point in the original number. Likewise there will be one digit to the right of the decimal point in the root for every pair of figures appearing in or added to the original number to the right of the decimal point. The last statement is particularly important in taking the square root of numbers less than one.

ACCURACY OF STATISTICAL DATA

The question of how many figures shall be retained in the result of a computation is particularly important in statistical work because many of the data employed are to some degree approximate. The problem of the statistician can be explained by contrasting his computations with those of the bookkeeper who is engaged in keeping a record of numerical facts in dollars and cents. The latter must keep his records accurate to two decimal places. Suppose the following inventory of raw materials was being prepared:

1,367 ft.	1 in. round iron at \$5.25 per 100 ft.	\$71.77
11,000 ft.	$2\frac{1}{2}$ in. $\times$ $\frac{1}{2}$ in. strap iron at \$7.62 $\frac{1}{2}$ per 100 ft.	\$838.75
etc.		

The first entry might be carried out to four decimal places, i.e., to \$71.7675, but the last two places are of no value to the bookkeeper who is interested in accuracy only to the nearest cent. Similarly the second entry is carried to cents although the 11,000 ft. may be only an estimate and even the \$838 not entirely accurate. The question of how many figures to retain does not arise in either of these cases nor does it arise in any case for the bookkeeper. The statistician is not in a similar position because more commonly he is dealing with data that are not expressed in dollars. Even when he deals with data expressed in dollars the question is not likely to be whether they should be carried to the nearest cent but whether to use a unit of \$100, \$1,000, or \$1,000,000.

Statistics is often defined as the science of large numbers, and properly interpreted this definition is sound, but to many it merely marks the statistician as one who works with figures containing six or eight

or even ten digits. Nothing could be farther from the facts than this impression. True, the statistician deals with aggregates of the magnitude of millions or billions, but a part of his working equipment consists in the use of well-established rules for rounding off large numbers.

Rounding Off Numbers

Meaning.—Precision work in a machine shop is seldom more accurate than to one part in a thousand. If the statistician dealing with data concerning the business world can achieve the same degree of accuracy as the machinist, the results will be amply satisfactory. Let us examine the meaning of data accurate to one part in a thousand. The average weekly earnings of factory workers in December, 1936, according to the National Industrial Conference Board, was \$26.63. This figure is an average obtained by dividing total weekly payrolls by number of workers employed and means that the result of the division was somewhere between \$26.625 and \$26.635. Hence a complete statement of the figure would be $\$26.63 \pm .005$. That is, a variation of as much as .005 may be present in 26.63, or a variation of 5 in 26,630 which is equivalent to 1 in 5,326. The average weekly earnings figure quoted to the nearest cent, therefore, is accurate to 1 part in 5,000 approximately.

From this example it will be clear that any figure quoted to four digits is accurate to at least 1 part in 2,000, on the assumption, of course, that the four quoted figures are accurate. Hence all the precision that is needed in statistical work can be provided by maintaining accuracy to four digits. In the preceding example this requirement was met by quoting weekly earnings to the nearest cent, but more generally four-digit accuracy will be sufficient regardless of the relation of the four digits to the position of the decimal point.

Significant Figures.—*In a single number or in the results of a computation the digits that show the extent to which the figure is accurate are called significant figures.* Some examples will help in understanding this definition. The number 98,000,000 has two significant figures unless it is known from the surrounding circumstances that the zeros are an accurate representation. If the actual amount represented may be anything between 97,500,000 and 98,500,000 then only the first two figures, 98, are significant and the zeros have no other purpose than to show the position of the decimal point. On the other hand if it is known that the actual amount lies somewhere between 97,999,-

500 and 98,000,500 then the original number would have five significant figures. That is, the first three zeros would be significant while the last three would serve the function of showing the position of the decimal point. Unless some indication to the contrary is given, the final zeros of a whole number are not to be considered significant. Likewise in a number less than one, zeros immediately following the decimal point are not significant. For example, .00042 has two significant figures, but .000420 has three significant figures because the final zero should be taken to mean that the operation was carried to three digits and the third one was found to be zero. That is, the actual facts are somewhere between .0004195 and .0004205. But .00042 means that the actual facts are somewhere between .000415 and .000425.

The argument of the preceding section can be summarized in terms of significant figures as follows: regardless of the absolute size of any datum⁷ not more than four significant figures need be retained for statistical purposes.

Method of Rounding Off.—When data are expressed to more than four significant figures, or more generally whenever a reduction in the number of significant figures is desired, methods of rounding off must be followed. There is no universally used set of rules for rounding numbers, but a set which has wide acceptance may be stated as follows:

1. When more than five is eliminated the preceding digit should be increased by one.
2. When less than five is eliminated the preceding digit should not be changed.
3. When exactly five is eliminated the preceding digit should be increased by one if it is an odd number but should not be changed if it is an even number.

Examples:

GIVEN NUMBER	ROUNDED NUMBER
1267862	1268
8762180	8762
5863500	5864
5862500	5862
5862517	5863

Sometimes a number rounded to four significant figures is subsequently rounded to three significant figures. This can be done by apply-

⁷ The word "data" should be used in a plural construction. The singular is "datum" referring to a single item or figure as used here.

ing the same rules, except for a number such as 467465. Rounded to four significant figures the number becomes 4675 and subsequently rounded to three significant figures according to Rule 3 it becomes 468, but obviously the result of rounding to three significant figures should be 467. This case is covered by an auxiliary rule sometimes followed by computers: "If a number when rounded upward ends in an even 5, indicate that fact by a prime (')." According to this rule the four-significant-figure result for the example would be written 4675' to indicate that in any subsequent rounding to three significant figures the third digit should not be increased to 8.

The foregoing is a statement of the mechanics of rounding off numbers. However, the statistician's problem does not end here, because in many cases the data with which he must work will not be accurate to four significant figures and usually the degree of accuracy is not stated. In such cases formal rules must be supplemented by a knowledge of the background of particular data. We proceed, therefore, to a detailed description of the kinds of figures which appear in statistical work and the basis for judging their accuracy.

Counting and Measurement

In statistical work enumerations of two kinds appear: (1) those in which the units are counted, and (2) those in which the units are measured. For example, the value of exports of 102 countries in 1935 as reported by the United States Department of Commerce was \$11,580,000,000. The number of countries included in the report was an exact count and any computations based upon it would not be subject to error. The value of exports was obtained by totaling the reports of customs of the several countries after converting the different monetary units to dollars, using some agreed upon set of exchange ratios. Due to inaccuracies of reporting within individual countries, variations in methods of valuing exports and the complication of applying exchange ratios between different monetary units, the figure for value of exports is at best only an approximate measurement.

Cases similar to both of these appear in statistical work. Units which are counted give rise to little or no difficulty in subsequent work. They may be accurate to five or six or more significant figures but not more than four need be retained in statistical work. On the other hand units which are measured immediately lead to the question:

How accurate are the results? Sometimes this question is answered specifically.

For example, the United States Department of Agriculture defines No. 1 Soft Red Winter Wheat as follows:

Minimum test weight per bushel.....	60 lbs.
(Scales accurate to one-tenth of a pound, hence the wheat must weigh more than 59.9 pounds to be No. 1 grade.)	
Maximum limits of	
Damaged kernels	2%
Foreign material	1%
Wheat of other classes.....	5%
(All of these are taken to the nearest per cent when tests are made.) ⁸	

Another example which states the margin of error is presented in Table 4.

TABLE 4

LONG-TERM PRIVATE DEBT: ESTIMATED AMOUNTS OUTSTANDING FOR 1912, BY CLASSES *
(Amount of debt in billions and tenths of billions of dollars)

CLASS OF DEBT	ESTIMATED DEBT	PER CENT DISTRIBUTION	MARGIN OF ERROR (per cent)
Total	31.3	100.0	10
Railway	10.7	34.2	..
Public utility	5.3	16.8	5
Industrial	4.5	14.4	10
Farm mortgage	3.8	12.2	4
Urban real estate.....	7.0	22.4	15

NOTE.—The margins of error shown in the table for private debt represent a non-statistical evaluation of the figures by the estimator.

* *Statistical Abstract*, 1936, United States Department of Commerce, Bureau of Foreign and Domestic Commerce, p. 275.

Examples of Measurement

More commonly the error to be expected is not indicated. Thus the user of the data is left to judge the degree of accuracy which can properly be attributed to them. Judgments of this sort must be based upon a knowledge of the method used in obtaining the data and a background of information concerning the source. For example, the United States Department of Agriculture announces in December the estimated crop of winter wheat for the year. The estimate as of December 1, 1937, was 873,993,000 bushels. The department receives annually about 160,000 reports from farmers in all sections of the

⁸ *Handbook of Official Grain Standards of the United States*, United States Department of Agriculture, Bureau of Agricultural Economics, revised June, 1937.

country; these include estimated acreage of wheat planted and estimated yield per acre. The two estimates are multiplied together to give approximate production in each locality. These approximate production figures are then weighted according to the probable total production which each represents, and combined. The result is an estimate of the production of wheat in the entire country. The process is actually much more refined than this incomplete statement would imply. The results obtained, while not entirely accurate, usually prove to be within 2 or 3 per cent of the actual production recorded in agricultural censuses.

An example of a different sort is the monthly report of floor space in new buildings contracted for, as compiled by the F. W. Dodge Corporation. The totals for thirty-seven states of the United States east of the Rocky Mountains are aggregates of the reports of local offices in all sections of this area, supplemented by reports from correspondents. The floor space in a building is estimated from the plans used in letting the contract for the building. These estimates are not intended to give the exact number of square feet of floor space; they may vary as much as 10 per cent from the actual area. When many such estimates are combined the underestimates tend to balance the overestimates so that the aggregate figure may be much more nearly correct than the individual estimates. However, in reporting non-residential construction contracts for January, 1938, as 9,637,000 square feet, an error of as much as 200,000 square feet might easily be present.

A third example is the report by the Bureau of Internal Revenue of the Treasury Department on net income of corporations for a year. The aggregate net income of all reporting corporations is carefully compiled from the bureau's files, hence the 1934 income of \$596,048,000 is considerably more accurate than the figures of either of the preceding examples.

These examples indicate the extent to which a background of knowledge of methods of collection is necessary in understanding the accuracy of data. The figures in the three illustrations are not equally accurate. The crop estimate may easily be in error by as much as 10 million bushels, hence not more than two significant figures of the estimate are accurate and it might as well be stated as 870 million bushels. There is false accuracy in stating this estimate to the nearest thousand bushels, because comparisons with the quinquennial census

of agriculture show that the estimates usually differ by several million bushels. False accuracy is common in published data but it causes little difficulty so long as the background of the data is sufficiently familiar for users to be aware that more significant figures have been retained than is warranted.

The same argument applies to the figure for floor space of construction contracts. It might better be stated as 9 million square feet,⁹ since a variation of as much as 200,000 square feet is inherent in the method of collecting the data. On the other hand the figure for corporation income tax is accurate to six significant figures, but there is no need to retain more than four significant figures; hence the figure should be written as 596.0 million dollars. The zero following the decimal point is written just as a digit other than zero would be written to show data accurate to the nearest hundred thousand dollars.

Significant Figures in Computation

The emphasis up to this point has been on the number of significant digits to retain in a single figure or a list of figures pertaining to a single subject. We are now ready to develop methods of dealing with rounded numbers in performing computations. The rules applicable to each of the four fundamental operations will be explained in order.¹⁰

In Addition.—Each of the examples in Table 5 illustrates a particular point in dealing with approximate numbers. In Example A exports from each division are given to the nearest hundred thousand dollars. This is done because the data are no more accurate than to that unit and because no greater accuracy is needed in statistical work. In the total there is no reason for retaining the fifth significant figure, and total value of exports for 1934 may be stated as 2,133 million dollars.

In Example B the operating revenues have been carried to the nearest dollar. The data are perfectly accurate since they come from audited statements submitted to the Interstate Commerce Commission by the

⁹ When the size of the unit in which data are expressed is increased the change should be from single units to thousands or millions or billions rather than to intermediate sized units. Thus 12,416,736 could be stated as 12,417 thousands if it were accurate to five digits, as 12.42 millions if it were accurate to four digits, and as 12.4 millions if it were accurate to three digits. The expression of the last two examples in the form 1,242 ten thousands and 124 hundred thousands, respectively, must be frowned upon because of the potential confusion in the minds of students as to the number of zeros to be added, if one wishes to return to the original unit.

¹⁰ These rules are not applicable to bookkeeping where accuracy must be maintained to the nearest cent regardless of the number of significant digits retained in a particular figure.

TABLE 5

EXAMPLES OF ROUNDING OFF MEASUREMENTS IN ADDITION

A		B	
VALUE OF UNITED STATES EXPORTS OF MERCHANDISE BY COAST AND BORDER DIVISIONS, 1934 *		OPERATING REVENUES OF CLASS I RAIL- ROADS OF THE UNITED STATES BY SOURCE, 1934 *	
DIVISION	VALUE OF EXPORTS (in millions and tenths of millions)	SOURCE	REVENUE
North Atlantic	\$ 810.8	Freight	\$2,629,301,525
South Atlantic	207.3	Passenger	345,889,550
Gulf Coast	509.9	Mail	91,139,847
Mexican Border	48.0	Express	54,013,025
Pacific Coast	259.8	Other	151,222,875
Northern Border	297.5	Total	\$3,271,566,822
Total	\$2,133.3	Rounded total	\$3,272,000,000
Rounded total	\$2,133.		

C

AREA OF LAND IN THE UNITED STATES FOR WHICH TITLE REMAINED WITH THE
GOVERNMENT ON JUNE 30, 1935 *

Use	NUMBER OF ACRES
National forests	138,710,942
National parks and monuments	8,724,737
Indian reservations (estimated net)	57,518,590
Military, naval, experimental reservations, etc. (approximate)	1,000,000
Unappropriated, but withdrawn (approximate)	197,261,754
Total	403,216,023
Rounded total	403,000,000

* *World Almanac*, 1936.

railroads. There is, however, no advantage in retaining ten significant figures in statistical work. According to the rule the operating revenue may be stated as 3,272 millions of dollars or the figure may be carried to dollars rounded off to the nearest million as shown in the table.

Example C differs from A and B in that part of the data are approximations. The area of the military, naval, and other reservations is estimated at 1,000,000 acres without any attempt to be more accurate. The areas of the Indian reservations and the unappropriated lands are likewise only approximate, yet the figures are given to an acre. There is false accuracy in these two figures, and an inconsistency in the table. *The result should not be carried beyond the limit of the least accurate figure* which appears to be millions of acres. The total, therefore, should be stated to only three significant figures.

The conclusion to be drawn from these examples is that usually no more than four significant figures are to be retained in a sum, and when the data are not accurate to four digits fewer should be retained to avoid introducing false accuracy in the result. It is not to be implied that all cases which arise will conform to these three examples. On

the other hand study of these examples will provide guidance in the selection of the proper number of significant figures to retain in any set of data.

In Subtraction.—The rules for subtraction are the same as those for addition. For example, a method which is commonly used in measuring the number of automobiles withdrawn from service each year includes these steps.

Number of automobiles registered, 1933.....	23,843,591
Number of automobiles produced for the domestic market, 1934	2,442,589
Number which could have been registered, 1934.....	26,286,180
Number actually registered, 1934.....	<u>24,933,403</u>
Number withdrawn from service, 1934.....	1,352,777

This method is not particularly accurate as a measure of cars "scrapped" because all second-hand cars taken in by dealers and not yet resold as well as cars which are temporarily unlicensed by their owners are included in the 1,352,777. At the moment, however, the rounding off of the figures is the point of interest. In spite of the fact that these figures imply an accurate count of automobiles, they are really subject to a substantial margin of error. The exact amount of error is unknown, but no inconvenience follows because there is no advantage in retaining more than four significant figures. The result would therefore be stated as 1,353,000 automobiles withdrawn from use during 1934. There is no certainty that these data are accurate even to the nearest thousand, but they would be assumed to be accurate to that extent unless definite information to the contrary were at hand.

In Multiplication.—The rule of four significant figures holds for multiplication just as in the preceding operations, but an additional rule must also be observed. *The product of two measurement numbers must not be retained to more significant figures than the least number of significant figures in either the multiplier or multiplicand.* For example, during May, 1936, 2,648,330 long tons of pig iron were produced in the United States. At that time the Composite Pig Iron Price was \$19.96 per long ton. The value of the month's production was $2,648,330 \times \$19.96 = \$52,860,666.80$. But the data for pig-iron production are approximate to an unknown extent and the composite price is an average which is accurate only to the nearest cent. Assuming that the production figure is accurate to the nearest hundred tons would give five significant figures, but there are only four significant figures in the price. Therefore, the value of production should be stated to

only four significant figures and could be written 52.86 million dollars. If the figure were written complete, it should be \$52,860,000. The reason for this approximation will be apparent from the two following computations which show the maximum and minimum values which this product may take when the last significant figure of each number is given its maximum and minimum value.

A	B
MINIMUM	MAXIMUM
2,648,250	2,648,350
19.955	19.965
<u>52,845,829</u>	<u>52,874,308</u>

These products differ in the fourth significant figure. It is therefore apparent that nothing beyond the fourth figure is of any value and indeed that the fourth figure is not exact, although sufficiently accurate for statistical work.¹¹

The rule for multiplying also includes the special case of squaring a measurement number. The significant digits of the square should not exceed the significant digits of the original number. For example, the square of the measurement 26.85 should be retained as 720.9.

In Division.—The rule for division is: *There should be no more significant figures in the quotient than the least number which appears in either the dividend or the divisor.* The general rule of not more than four significant figures in a statistical calculation also applies. For example, the December 1, 1936, final estimate of the United States cotton crop for 1936 according to the Department of Agriculture was 12,399,000 bales. The estimate of acreage harvested was 30,028,000 acres. The average yield per acre is obtained by dividing the production by the acreage, i.e., $12,399,000 \div 30,028,000 = .41291$ bales per acre. The result may be carried to five digits according to the rule that the significant figures retained in the quotient should not exceed the number of significant figures in either dividend or divisor, whichever is smaller. However, statistical work usually requires keeping only four significant figures. Rounding off to four places, then, the production is .4129 bales per acre. Actually the result would be expressed in pounds by multiplying the average in bales (.4129) by 500. The average yield in pounds would be 206.5 per acre.

¹¹ George G. Chambers, in *An Introduction to Statistical Analysis* (F. S. Crofts and Co., New York, 1925), would not retain the fourth significant figure. His rule is "if the product of two single number approximations is expressed as a single number approximation the integer [significant figures] of the product is less than the integer [significant figures] of the least accurate factor," p. 27.

Suppose that the acreage harvested were not considered accurate to the nearest 1,000 acres, but were reported as 30.0 million acres accurate only to the nearest 100,000 acres. The average yield per acre would then be $12,399,000 \div 30,000,000 = .413$ bales. The result can be carried to no more than three significant figures because only three are significant in the divisor. The reason for retaining only three figures will be apparent, if the divisions are made giving dividend and divisor their minimum and maximum values.

A MINIMUM	B MAXIMUM
$\frac{12,398,500}{30,050,000} = .4126$	$\frac{12,399,500}{29,950,000} = .4140$

The fourth figure has no significance whatever since the two quotients do not agree even in the third figure. As stated in the discussion of multiplication, the third figure is somewhat approximate but is accurate enough for use in statistical work.¹²

In Extracting Square Roots.—The reverse of the rule for squaring numbers holds for square roots. That is, *as many significant figures may be retained in the root of a measurement number as there are in the number.* Hence $\sqrt{327} = 18.1$.

SUMMARY

The purpose of this chapter was stated as an attempt to set forth in elementary form a background of computation methods which would facilitate the work of subsequent chapters. The first part is devoted to a review of arithmetic processes while the second deals with the rounding off of figures for statistical purposes and the rules for determining the number of significant figures to be retained. The material presented is, of course, germane to all operations with numbers, but is particularly useful in statistical computation.

The primary task of the statistician is not, however, to make of himself a "figuring fool." The most important task is mastery of the techniques which will be developed in the chapters which follow. Ability to compute rapidly and accurately must be considered as the necessary background for, but not the main object of, statistical work.

¹² Perhaps attention should be directed again to the meaning of the expression "good enough for statistical work." The figure .413 bales per acre appearing in print should be taken to mean not less than .4125 and not more than .4135. The variation in either direction is .0005 on .413 or 5 on 4130 which is a variation of one part in 826 and this is accurate enough for statistical work.

PROBLEMS

Problems 1-10 are self-tests in which the student can check his own performance against the standard time listed. Do not write answers in the book in order that additional trials can be made if the first one fails to meet the time limit.

1. Addition (40 seconds)

943	167	4956	2286	6269
376	742	6237	7463	4728
641	969	312	9498	8247
879	378	8468	4537	3722
<u>2674</u>	<u>2776</u>	<u>9112</u>	<u>3786</u>	<u>1122</u>

2. Addition (3 minutes)

876	24.29	476,876	31.35	.4832	1371.10
937	15.15	377,139	42.50	.1887	1229.48
711	41.69	991,387	1.46	.0942	782.20
492	39.63	398,872	23.59	.3948	59.35
321	23.15	814,612	5.62	.0038	2892.36
173	15.12	329,388	7.35	.1850	755.73
288	4.28	376,441	66.75	.3763	3842.45
317	16.14	114,473	103.43	.0382	4721.21
222	34.99	787,224	35.78	.0976	2783.29
384	55.29	716,326	7.11	.1956	1972.48
<u>244</u>	<u>24.6</u>	<u>444444</u>	<u>244</u>	<u>244</u>	<u>6244</u>

3. Subtraction (35 seconds)

1090	8617	31.762	217.32	\$27,218.45	\$586.89
<u>-585</u>	<u>-7758</u>	<u>-4.86</u>	<u>-29.685</u>	<u>-11,216.10</u>	<u>-497.98</u>

4. Multiplication (1 minute, 30 seconds)

921	875	726	486	1269	8296
<u>23</u>	<u>19</u>	<u>68</u>	<u>35</u>	<u>137</u>	<u>864</u>

5. Multiplication (2 minutes, 30 seconds)

3.8125	34.4167	.2976	620.14
<u>8.875</u>	<u>21.72</u>	<u>1.093</u>	<u>7.963</u>

6. Division (1 minute, 15 seconds)

237) 50481	593) 28464	418) 240768
------------	------------	-------------

7. Division (carry to four significant figures) (3 minutes, 30 seconds)

29.57) 128.43	.2448) 107.321	224.08) 3.11417
---------------	----------------	-----------------

8. Multiplication by short-cut methods (1 minute, 15 seconds)

793 × 25	65 × 65
2641 × 33½	93 × 107
732 × 199	47 × 47

9. Multiplication by short-cut methods (2 minutes)

$$2.183 \times .875$$

$$48027 \times 9\ 01$$

$$115 \times 115$$

$$892 \times 908$$

$$81 \times 81$$

$$176 \times 176$$

10. Division by short-cut methods (1 minute)

$$5418212 \div 25$$

$$83.47 \div .05$$

$$.4983 \div 12.5$$

11. Find the value of

$$a) (28 \times 37) + (12 \times 16) - (31 \times 29)$$

$$b) 3 + (9 \times 36) - (22 \times 11) + 486 + (138 \div 6)$$

$$c) 1417 - 12\{16 + 9(21 - 8)\}$$

$$d) (86 \times 22) \div 44 + (98 \div 7) \times 210 - 432 \div (12 \times 12)$$

$$e) 1217[81 \times \{5952 \div (31 \times 32) + 4\} - 900]$$

12. Find the value of

$$a) \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16}$$

$$b) \frac{1}{3} - \frac{1}{5} + \frac{1}{7}$$

$$c) 3\frac{2}{9} + 10\frac{4}{15} - 6\frac{1}{2}\frac{1}{4}$$

13. A can do a piece of work in 15 days, B can do the work in 18 days, and C in 25 days. What fraction of the work will the three working together perform in six days?

14. In Problem 13, after A has worked four days and C six days, what is the difference in the fraction of the work performed by the two?

$$15. 12\frac{3}{10} \times 105\frac{5}{9} \div 649\frac{1}{8} = ?$$

16. Mr. Smith invested \$27,500 in a partnership having a total capital of \$200,000. He later sold $\frac{1}{3}$ of his holding in the concern. What fraction of the ownership of the concern did he retain? What fraction did he sell? What amount should Mr. Smith receive of a \$12,000 profit, (1) originally, (2) after selling $\frac{1}{3}$ of his equity?

17. If 16 items are to be plotted at equal distances and centered on a sheet of graph paper $9\frac{1}{2}$ inches wide and the space allotted to each item must be a multiple of $\frac{1}{4}$ inch, how much space will be left for margins at each side of the paper?

18. If $14\frac{3}{8}$ tons of coal cost \$122 $\frac{3}{16}$ what was the cost per ton?

19. Express the following as (a) decimals, (b) per cents: $\frac{1}{30}$, $\frac{4}{9}$, $\frac{1}{16}$, $\frac{7}{16}$, $3\frac{3}{8}$, $13\frac{3}{16}$.

20. Express the following as (a) common fractions, (b) per cents: .06, .003, .004167, .65, 3.1875.

21. Express the following as (a) common fractions, (b) decimals: 10 per cent, 6 per cent, $18\frac{3}{4}$ per cent, $\frac{2}{3}$ per cent, $262\frac{1}{2}$ per cent.
22. Arrange each of the following in ascending order of value:
 (a) .43, $\frac{5}{12}$, 37.5 per cent, .4, $\frac{21}{50}$
 (b) $\frac{1}{2}$ per cent, .086, $\frac{2}{25}$, $\frac{3}{20}$ per cent, $1\frac{7}{100}$.
23. The spoilage on two crates of oranges each containing 210 oranges was 17 per cent and 33 per cent, respectively. Find the income from sale of the unspoiled oranges at 45 cents per dozen.
24. Assessments in a city are maintained at 70 per cent of market value. If Mr. White pays \$302.40 tax when the tax rate is \$30 per thousand of assessed valuation, what is the market value of his property?
25. The balance sheet of a concern showed the following:

Cash	\$ 7,500
Accounts receivable	38,500
Inventory	15,750
Investments	9,050
Plant	120,000
Equipment	69,200
Total assets	\$260,000

 Each type of asset was what per cent of the total?
26. An article cost \$450. At what price should it be sold (a) to make a profit of 55 per cent on cost, (b) to make a profit of 45 per cent on the selling price?
27. If a worker's wages are cut 25 per cent and subsequently increased 25 per cent, the most recent wage is what per cent of the original wage?
28. Given the following:

MONTH	GROCERY SALES	NO. OF DAYS STORE WAS OPEN
July	\$28,412	24
August	29,827	26

Find the per cent of change in average daily sales in August compared with July.

29. The following information is available concerning the manufacture of a particular article:

	1939	1940
No. of units produced.....	200,000	275,000
Overhead costs	\$50,000	\$50,000
Variable costs	\$100,000	\$120,000
Sales income	\$200,000	\$275,000

- a) The per cent of profit on selling price increased by what per cent in 1940?

- d) The total assets of all member banks of the Federal Reserve System on the December 31, 1937, call date were \$46,785,512,000.
- e) The Bureau of Foreign and Domestic Commerce of the Department of Commerce estimates from a sample collection that the total retail trade of the United States in 1937 amounted to \$39,930,000,000.

In each of the following problems express the summary figures to the correct number of significant digits.

36.

A

LIABILITIES OF FEDERAL INTERMEDIATE CREDIT BANKS, DECEMBER 31, 1937

LIABILITY	VALUE (thousands of dollars)
Paid in capital and surplus, United States government.....	100,000
Surplus, earned reserves and undivided profits*.....	12,561
Debentures outstanding (unmatured)†.....	174,950
Total	287,511

*Net amount after deducting impairment or deficit.

†Adjusted for debentures held by banks of issue and by other federal intermediate credit banks.

B

PRODUCTION, TRADE, AND SUPPLY AVAILABLE FOR CONSUMPTION OF RAW SUGAR, CONTINENTAL UNITED STATES, 1935

ITEM	QUANTITY (short tons)
Production (beet and cane only).....	1,651,000
Brought in from insular areas.....	2,686,969
Imports as sugar.....	2,372,066
Exports as sugar.....	103,349
Exports in other forms.....	13,220
Available for consumption.....	6,593,466

37. (a) In the following table, how many significant figures should be retained in the total consumption? (b) Assuming the accuracy of a population of 129,257,000 in 1937, what is the per capita consumption of meats in the United States?

PRODUCTION, FOREIGN TRADE AND CONSUMPTION OF ALL MEATS IN THE UNITED STATES, 1937

ITEM	AMOUNT (million pounds)
Production	
Federally inspected	10,273
Uninspected (estimated)	5,299
Exports of United States production.....	164
Imports for consumption.....	263
Net change in storage stocks, decrease.....	402
Consumption	16,073

38. Find the value of a corn crop estimated at 4,000 bushels, if it was sold for $87\frac{1}{2}$ cents per bushel.
39. A motorist drove 3,532 miles from Boston to San Francisco, using $207\frac{1}{2}$ gallons of gasoline. What was his average mileage per gallon?

REFERENCES

CHAMBERS, GEORGE G., *An Introduction to Statistical Analysis*. New York: F. S. Crofts and Co., 1925.

Chapters I and II deal with measurement, approximation and significant digits.

EDGERTON, EDWARD I., and BARTHOLOMEW, WALLACE E., *Business Mathematics*. New York: The Ronald Press Co., 1923.

Chapter XII explains short methods of computation and various checks of accuracy.

LANGER, CHARLES H., and GILL, T. B., *Mathematics of Accounting and Finance*. Chicago: Walton School of Commerce, 1930.

The first five chapters contain a detailed statement of the fundamentals of arithmetic and algebra. Short cuts are presented in pp. 43-64.

MURPHY, PATRICK, *Short Practical Rules for Commercial Calculations*. Albany: Weed-Parsons Printing Co., 1910 (originally printed in 1886).

A highly stimulating presentation of short-cut methods for the student who cares to pursue the subject at length.

WALKER, HELEN M., *Mathematics Essential for Elementary Statistics*. New York: Henry Holt and Co., 1934.

Chapters I-VI contain material similar to part of the text. Chapter II on significant figures is particularly pertinent.

CHAPTER III

STATISTICAL INVESTIGATION

THE CHARACTER OF STATISTICAL INVESTIGATION

THE EXTENT to which the work of the statistician underlies the conduct of business affairs was discussed in chapter I. Sometimes the contribution which he makes is relatively simple, being confined merely to presenting sales figures graphically. On the other hand his task may consist in a study of sales records and indexes of regional purchasing power for the purpose of determining sales territories and establishing sales quotas, or the study of a sample of output to determine whether it meets contract specifications. Whatever the complexity of a particular problem, the sequence of steps followed in its solution involves the application of statistical method.

Definition of Statistical Method

The statistical method is essentially the use of the principles of scientific investigation in the study of aggregates of numerical information. Just as the physicist must develop laboratory methods and techniques for examining the theories of sound, light, etc., so the statistician must have methods of appraising the theories of probability and sampling in terms of the observed phenomena (numerical data) of the business world. The problem of the statistician is complicated considerably by the fact that business operations cannot be subjected to the control that is possible in the physics laboratory. As a result the methods of statistical investigation are those research procedures developed to meet the peculiar requirements of the problems arising in the conduct of business affairs.

An example will demonstrate the difference between the controlled conditions of the physics laboratory and the uncontrolled conditions of the statistics "laboratory." The physicist wishing to read the height of a column of mercury in a manometer tube sets up his apparatus, provides for a constant temperature in his laboratory, selects a time at which barometric pressure is stable, and proceeds to take a large number of readings on the scale attached to his apparatus. The aver-

age of a large number of such readings will be the theoretically best value for the height of the column of mercury in the tube. In contrast to this, suppose that a statistician wishes to determine the per cent of the employable workers of the United States who are unemployed as of a given date. He might also elect to take a large number of independent readings of the phenomenon under investigation and take the average of the results as the best value for the per cent of workers unemployed. But he encounters a whole mass of preliminary problems before any observations can be made and none of these can be "controlled." He must define "unemployed person," "employed person," "employable person," and no matter how carefully these definitions are phrased doubtful cases will arise. He must determine how to select samples of the population which will be representative, and even the most meticulous care will not produce a result comparable with the stability of the column of mercury in the physicist's laboratory. These and similar problems have forced the statistician to develop methods of investigation which are peculiar to the type of data with which he deals and the uncontrolled conditions under which he must use them.

The employment of statistical methods in the solution of business problems belongs almost exclusively to the twentieth century. At an earlier date when business enterprises were small, management was able to comprehend its problems in detail by personal contact. The increased size of concerns in the present period has required more planning and greater regimentation of operations. At the same time management has found it impossible to maintain personal contact with its problems. The alternative is control through the interpretation of numerical information. This chain of circumstances has led to the introduction of statistical methods of investigation as a primary aid in the performance of the function of management.

The Use of Statistical Method

Masses of Data.—The methods used by life insurance actuaries give no information concerning the time at which a particular insured person will die, but they give very accurate information concerning the number of persons who are likely to die in any year out of a large number of a given age alive at the beginning of that year. Life insurance premiums are based upon the regularity of death rates among large groups of persons, not on a guess as to how long an individual will survive. Similarly, a study of department-store experience

may show that bad debt losses on charge accounts amount on the average to about 1 per cent of charge sales. It does not follow that an individual store must have bad debt losses of 1 per cent. This result can be applied to particular cases only by taking account of the relation of conditions in the individual case to the average conditions found in the large group. The individual store may have 2 per cent bad debt losses in a certain year due to the fact that its customers have been experiencing the effects of a great amount of unemployment. In another year when its customers are fully employed the same store may have only $\frac{1}{2}$ of 1 per cent bad debt losses. Another example is the use of income tax statistics in the determination of sales quotas. Studies show that the higher the percentage of the population of a state filing income tax returns, the higher the percentage of the population purchasing automobiles. This relationship can be used to establish sales quotas for automobile agencies in the various states. It does not follow, of course, that those persons who file income tax reports will necessarily purchase automobiles, but that the higher purchasing power evidenced by the larger percentage filing tax reports will be available for the purchase of automobiles. Hence intensified sales effort where the purchasing power exists should produce the best sales results.

These three examples show how management uses the results obtained from the study of mass information. The typical situation found in the group is used as a guide for action within individual concerns.

Case Investigations.—There is, however, one type of statistical work known as the case method which does not deal with masses of data. An individual case is studied intensively, usually over a period of time, in order to make a complete analysis of its operations. The *case* may be one individual, a single family, a business concern, or any other similar entity.

In statistical work case investigations are of less frequent occurrence than mass data investigations. More often than not case studies eschew statistical method entirely and rely solely on historical description in the presentation of results. A case study is characterized by the establishing of such a strong personal relation between the investigator and the person or persons furnishing information that a vast amount of detailed information can be obtained concerning the case. In presenting the results, each case is written up separately and represents a

complete investigation in itself. The distinguishing feature of the case method is the fact that a detailed description of the individual case is the objective.

The records maintained by physicians concerning their patients become most complete life-histories, sometimes covering the entire span from the cradle to the grave. These records are, of course, confidential, but their anonymous publication would provide a remarkable background for the study of sickness and health problems. In a similar fashion a file for a period of years of a financial manual such as *Moody's Manual of Investments*, giving as it does a brief case history of many individual corporations, becomes a compendium of invaluable case records of the founding, growth, financial organization, and in some instances the decline of individual concerns. These are available for study either as individual cases or collectively as the raw material for statistical analysis.

Case study is used infrequently in the investigation of business problems. It is a method that is well adapted to studies of social phenomena and has been widely used in the field of social work. As such it lies outside the scope of this book.

THE CANONS OF STATISTICAL INVESTIGATION

The attitude of the statistician toward his work is a matter of considerable importance. His methods are equivalent in the field of social science to those employed in the exact sciences by the chemist, physicist, and biologist and his attitude toward his work must be equally scientific. Under no circumstances can he become an advocate or a special pleader. Statistical work done for purposes of pleading does not deserve the name of scientific research.

As a means of promoting the scientific character of statistical investigation there are certain standards or requirements which should be uniformly maintained. These fall naturally under three heads, each of which requires detailed explanation.

Definite Object

Statistical investigation is never aimless. It is always directed to the solution of a specific problem. The problem may be as basic as finding the total annual income of the nation or as circumscribed as a study of the amount of flour hauled on the New York State Barge

Canal during September, 1937. But regardless of scope the purpose must be specifically defined. Unless this requirement is met, direction will be lacking in the investigation, unnecessary work will be done and results of questionable value will be obtained. It is essential therefore to have the exact object of an investigation fully understood before any other work in connection with it is undertaken. At all subsequent stages of the investigation the purpose must be kept in mind as a guide in the planning and execution of the project.

Unbiased Attitude

The statistical investigator sets out to determine by investigation the facts concerning a given problem, but not to prove a certain thesis. There are times when it is very difficult to maintain an unbiased attitude. Some questions are of such controversial character that even the most detached investigator finds himself influenced. On the other hand in reading the report of an investigation one frequently has a feeling that the author has "leaned backward" to avoid bias. This is the proper attitude of a careful investigator when he finds himself placed at the center of a controversy.

Conscious or unconscious bias may appear in statistical work. Conscious bias can be dismissed quickly. A person who willingly distorts statistics for the purpose of proving a preconceived idea should not be called a statistician. He is a propagandist. It is necessary to be vigilant at all times to avoid using results containing bias. Conscious bias may appear in one or several of the following forms: (1) direct misstatement, (2) ambiguous statement, (3) the use of only favorable data, (4) concealed shifting of units of measurement, (5) deliberate selection of incorrect techniques, and (6) misleading forms of presentation.

Careful study is usually required to detect unconscious bias. Perhaps it would be safe to assume that all statistical interpretations contain some bias but that in most cases it is not present to a harmful degree. This is only another way of saying that the results of statistical work must be interpreted by human beings, each of whom can interpret only in terms of his own experience and his attitude toward the problem at hand. An excellent example of unconscious bias appears in the writings of certain statisticians and economists who during 1928 and 1929 interpreted the trends of the times to mean that permanent prosperity at the then existing levels had been attained. Subsequent events have shown that these men were so enamored of the favorable factors that

they overlooked the growing stresses in our economic system. Their biased attitude is apparent now, but at the time their teachings had a wide acceptance.

Skepticism

The beginner in statistical work is likely to have the attitude that numerical facts can be accepted without question. A few adverse experiences will usually dispel this initial trustfulness. The attitude of faith should then be replaced by skepticism or in the extreme by cynicism, because it is far better to err in that direction than to develop enthusiasm with its attendant misinterpretation. Many of the fallacies which appear in statistical presentation arise from failure of those responsible for the results to maintain a critical attitude toward their work.

STEPS IN STATISTICAL INVESTIGATION

As stated at the beginning of the chapter, there is a logical sequence of steps to be followed in statistical investigation. An outline of these steps will give the reader a view of the process as a whole prior to studying the details.

- I. Statement of the problem
- II. Preliminary planning of the investigation
- III. Collection of data
 - A. Library sources
 - B. Direct sources
- IV. Analysis of data
 - A. Editing of collected information
 - B. Tabulation
 - C. Ratios
 - D. Graphs
 - E. Measures of central tendency
 - F. Measures of dispersion and skewness
 - G. Index numbers
 - H. Time series analysis and application
 - I. Correlation and variance
 - J. Tests of reliability of samples
- V. Interpretation and application of the results of analysis
- VI. Preparation of a report of the completed investigation

The remainder of the book is devoted to a detailed presentation of the work involved in following through the several steps. Although the emphasis given on subsequent pages to the different parts of this outline depends upon the difficulty and ramifications of the particular subject, it is to be hoped that the reader will not lose sight of the fact that with one necessary exception he is following the outline step by step. The interpretation of the results of each type of analysis is not a procedure that can be relegated to a separate section of the book. Therefore the discussion of illustrative examples has been woven into the text wherever it has seemed desirable.

THE SCOPE OF DIFFERENT INVESTIGATIONS

The six major steps presented here cover the complete sequence of things that must be done in conducting an investigation, no matter how limited or how broad its scope. The subheadings are partly alternative and partly sequential depending on the character of a particular problem. The amount of detailed planning required and the time consumed in executing the plan will, of course, vary with the size and importance of the investigation. The type of planning in turn is directly related to the question of whether the investigation is internal or external in character. An internal investigation is one which deals exclusively with conditions within a single business concern or agency. Those investigations that originate outside the management of any particular business concern are called external. There is one great difference between internal and external investigations: the former as a rule present no serious problems of collecting data, whereas the latter are seldom free from such problems.

Internal Investigations

Statistical studies by a business concern of its own records are usually conducted to obtain information needed to assist management. The most common examples are found in the work of the cost accounting department. The data for determining unit costs are found in the accounting department and in the plant production records. The task of allocating the various factory and overhead costs to obtain an average cost per unit of product requires the use of statistical techniques. The cost accountant ordinarily does not employ all of the steps of investigation as outlined. His task is a rather circumscribed one, but it

is necessary for him to be familiar with the complete process in order to make intelligent use of the part that he needs.

Internal investigations of a more general character are a necessary part of business control, and these are likely to make use of more of the steps of statistical investigation. For example, an oil company wishing to study the weekly rhythm in the sales of gasoline at its filling stations in different parts of a city would have no difficulty in collecting data, since that information would be included in the daily report of the manager of each station. Combining the sales reports for a number of weeks to obtain an average relationship would involve certain adjustments for weather conditions, for any irregularities at the station which might affect sales, for holidays and other circumstances of a similar nature. Once the pattern of the weekly rhythm at each station had been obtained, the next step would be to study these patterns in pairs and groups to discover in what parts of the city similar rhythms appeared. This information could be used by the central office in assigning attendants so as to provide the maximum service to customers, in planning delivery schedules of tank trucks, as well as in planning the location of new stations.

In this example the emphasis is on analysis and interpretation, and that will be found commonly true of internal investigations. Although relatively simple statistical techniques are involved in this case, trustworthy conclusions depend upon following the steps of investigation faithfully. This leads to the general observation that a knowledge of the steps of statistical investigation and the relation of each to the whole is necessary to protect the statistician from error, even though his particular problem involves the use of only a part of the whole procedure.

External Investigations

Investigations conducted by manufacturers' and trade associations, advertising agencies, research bodies, universities, and government agencies are ordinarily more general in character than internal investigations. Correspondingly they call for the use of a wider range of statistical techniques. In particular, the preliminary planning and the collection of data demand much more attention in external investigations. For example, the plans for the 1940 census of population were under way as early as 1937 in the Census Bureau at Washington and in the field with co-operating agencies. The field work required only

a few weeks early in 1940, but a large staff will be continuously engaged for the next decade in preparing and publishing the various tabulations and analyses of the collected information. The entire process, in reality a continuous statistical investigation of the population, falls within the framework of method outlined in the preceding section.

Sometimes the scope of an investigation is limited in the sense that not all of the successive steps are carried on by a single agency. The task may be confined to collecting data. If so, the steps following collection can be ignored, but those preceding actual collection must be given proper attention. Again the particular task may be confined to interpretation and presentation of data collected by someone else, but the preceding steps must be thoroughly understood before any attempt is made to explain the meaning of the results.

SUMMARY

All of this discussion points to the same conclusion, namely, *no matter how simple a particular piece of statistical work may be, its execution requires a knowledge of the steps in statistical investigation.* Through this principle the various details of statistical method and technique are welded into a unified whole. Succeeding chapters are arranged so that the steps of statistical investigation will appear in natural sequence. Within this sequence the methods of analysis progress from those using simple techniques to those which are more involved.

PROBLEMS

1. What are the differences between research in the natural sciences and statistical research?
2. Describe an example from your own experience of the use of mass data in statistical work.
3. State a definite subject for investigation in each of the following fields:
(a) automobiles, (b) cost of living, (c) athletics, (d) profit.
4. What changes would you suggest in the conclusions reached in each of the following examples?
 - a) Hourly wage rates in industry have increased uninterruptedly for the past 20 years, and the cost of living is lower today than it was 20 years ago. Therefore the living standard is higher today than it has ever been in the past.

- b) In the District of Columbia in a recent year one male automobile driver in every 1,370 was involved in a fatal accident and one female driver in every 9,090 was involved in a fatal accident. Therefore women are safer drivers than men.
 - c) During World War I the United States army lost 126,000 men killed in action, died of wounds, and died from disease or other cause out of 4,355,000 men mobilized, a death rate of 28.9 per 1,000. During the years 1917 and 1918 the death rate for the United States exclusive of the armed forces was 32.4 per 1,000. Therefore it was safer in the army than at home.
5. Explain the differences between internal and external investigations.
 6. Which of the steps of statistical investigation were employed in preparing the reports appearing as Examples 1, 2, and 3 in chapter I, pages 7-9? Give references in the examples of specific statements which indicate the use of the several steps named in your answers.

REFERENCES

CROXTON, FREDERICK E., and COWDEN, DUDLEY J., *Applied General Statistics*. New York: Prentice-Hall, Inc., 1939.

Chapter I emphasizes the parts of statistical investigation and the kinds of errors that appear in statistical work.

MILLS, FREDERICK C., "On Measurement in Economics," *The Trend of Economics*. New York: Alfred A. Knopf, 1924, pp. 37-72.

An advanced statement of the place of statistical investigation in the realm of science.

SECRIST, HORACE, "Statistical Standards in Business Research," *Quarterly Publications of the American Statistical Association*, Vol. XVII, No. 129 (March, 1920), pp. 45-58.

An article that helped to establish standards in a period when business research was less firmly established than at present.

SPAHR, WALTER E., and SWENSON, RINEHART J., *Methods and Status of Scientific Research*. New York: Harper and Bros., 1930.

Every embryonic statistical investigator should be familiar with the point of view expressed in chapters I-VI.

CHAPTER IV

PRELIMINARY PLANNING OF INVESTIGATIONS

INTRODUCTION

INEXPERIENCED collectors of data sometimes make the mistake of jumping directly into the task of collection without an adequate comprehension of the problem with which they are dealing. This practice should never be followed no matter how simple and direct the problem may appear to be. There are always preliminary points which should receive attention prior to the actual collection of data. The four major steps which should be followed are: (1) define the problem; (2) study the problem; (3) plan the procedure; (4) prepare a statement of the program.

DEFINE THE PROBLEM

At the outset a crude statement will serve as a focus for the initial consideration of what is involved in the problem, but the crystallization of a few ideas will very quickly provide a mental setting and point to some of the limitations which should be established as a basis for more careful planning of the investigation. These preliminary ideas should be brought together in a more complete definition which will indicate the subject to be investigated, the exact object of the investigation and the limitations upon its scope.

An example will demonstrate the difference between an incomplete and a complete statement of the subject for research. Suppose that the statistician were to receive the following problem: "The sales of our company declined last month. This decline was unexpected, since all parts of the organization appeared to be unusually busy. Investigate the matter." This statement does not define the problem for research. It could be taken to mean that an investigation was wanted of why the organization appeared to be unusually busy; but assuming that an investigation is wanted of why sales declined, the statistician needs more information before proceeding with the work. Questions such as the following must be settled:

Are all products to be included or only major ones? If the latter, which products?

Is the investigation to be confined to discovering the facts or shall it include data pertinent to discovering the cause of the decline?

How much time is available for making the investigation?

Have the affected departments agreed to co-operate?

With questions of this sort settled, the problem might be restated somewhat as follows: "Investigate the extent of the decline last month in the sales of the five major products which we manufacture and present as much collateral information as possible to aid in determining the cause of the decline. Your report should be available prior to the directors' meeting which will be held three weeks hence." The purpose of the investigation is clear and the limitations as to time and scope are definite.

This example illustrates the type of definition required in an internal investigation to be carried out within a short period. Larger problems will require correspondingly greater amount of definition.

STUDY THE PROBLEM

Read About the Problem

¹ A knowledge of previous work that may have been done on a problem should be acquired as background before a new investigation is undertaken. The existence of earlier studies can be determined primarily through a search of library files. One may find that the problem has been investigated previously and that any further investigation should be built upon the existing work. Again flaws may be found in the previous work which make it completely or partially useless for the purposes in view. The chief value of studying such previous investigations may lie in discovering what *not* to do.

Library search may disclose the fact that no similar statistical investigation has been made previously. But books and magazine articles may be discovered which give factual information dealing with some phase of the subject or clues concerning methods of investigation. Library reading on the subject will aid the investigator in avoiding duplication of work already done; in avoiding the errors made in previous investigations; in discovering methods of approach and procedure; and in acquiring a broad perspective of his problem.

Finally there will be some cases in which no usable information of

any sort can be gleaned from library search. When this happens the investigator must be prepared to proceed without such assistance. He must be able to supply from his own experience the background that otherwise would have come from library reading.

Think the Problem Through

At this stage the investigator should take some time for thoughtful consideration of his problem. There are major parts of his plan which should be settled. Certain parts may need additional emphasis and others should perhaps be discarded. New phases may enter as a result of the reading done. The knowledge which has been acquired by reading needs to be related to the particular problem at hand. The investigator should be able to visualize his entire procedure. At first this should be confined to the main outline of the work, and following that the details should be considered. In constructing this mental image of the work it is unwise to assume that the preliminary planning can ignore details that are apparently simple, for they may contain difficulties. The success of the true investigator lies in his ability to foresee these concealed difficulties and make provision for them. The case of the student who wrote to a number of cement companies asking for production in tons and price per barrel of cement will illustrate the point. The companies had to change their figures, which were recorded in barrels, to tons to meet the student's questions and the student in turn had to change from tons back to barrels when he tabulated the data. A small amount of foresight would have avoided the difficulty.

A word of caution may aid in avoiding misinterpretation of the preceding paragraph. While the plan of procedure should be thought out very carefully, it is scarcely to be expected that no subsequent changes will be necessary. Regardless of how efficient the investigator may be, it is unlikely that he can foresee and provide for every contingency which may arise. The plan should therefore be sufficiently flexible to permit necessary adjustments to conditions as they develop.

PLAN THE PROCEDURE

The amount of planning needed will be determined by the complexity of the problem and the size of the investigation. In some cases the points discussed in this section will take care of themselves, but

more commonly decisions concerning them must be made prior to beginning the collection of data. Under either circumstance consideration must be given to the elements of the plan lest some essential be overlooked.

Library Sources and Direct Sources

The reading which has been done on a problem should indicate fairly well whether the needed data will be available in libraries or whether recourse must be had to direct sources. It will not serve merely to remember that some data on the subject were referred to in a book or magazine article. The data must be found and examined. Then several questions must be settled. Are the data in usable form? Do they include the desired time period? Do they cover the proper area, i.e., nation, state, locality, etc? Are they expressed in the correct unit for the particular purpose? Are they reliable? By the time these questions have been investigated the general problem of whether or not library sources can be used will have been definitely settled. If library sources can be used, the investigator is ready to move on to the next part of his plan.¹ If on the contrary library sources should fail to provide any or all of the required data, the possibility of securing them directly must be canvassed.

Using direct sources² means going to the business concerns, agencies, or individuals possessing the information to obtain at first hand data which do not exist anywhere in print. In some cases the preliminary survey may disclose the fact that the desired data cannot be found in library sources and that they are equally unavailable from direct sources.

¹ The details concerning the collection of data from library sources are presented in chapters IX and X.

² The classification of data as library and direct, a distinction according to source, represents something of a departure from the usual classification found in textbooks. The customary division into primary and secondary data places the emphasis on the number of times that data have been recorded, i.e., primary data are those which are being recorded for the first time by the investigator who assembles them, whereas any subsequent recording of the data by other than the original investigator makes them secondary data. For example, the report of steel production found in the *Annual Statistical Report* of the American Iron and Steel Institute is primary data and the *Report* is a primary source, whereas the same figures published in the *Survey of Current Business* of the United States Department of Commerce become secondary data, and the *Survey* a secondary source. The names, primary and secondary, suggest that the former are more reliable than the latter. The point of view of this book is that reliability does not depend so much upon the number of times the data have been handled, as upon factors related to the canons of statistical investigation discussed in the preceding chapter—factors which are as likely to affect one kind of source as another.

The distinction between library and direct sources places the emphasis on methods of procedure. One type of research is required to obtain data already available for general use, but quite a different type of work is required to obtain data directly from the originating source.

For example, recently the Minimum Wage Division of the New York State Department of Labor undertook to obtain data on hours worked daily and weekly by operators in beauty parlors. The nature of the work in this trade is such that hours actually worked vary greatly from stated schedules, but are unrecorded except in the very large establishments. As a result, the field workers found it very difficult to secure accurate information. When obstacles of this sort are discovered, the choice lies between abandoning at least those difficult features of the investigation, or continuing with the understanding that the results will have only conditional validity. On the other hand if the preliminary survey indicates that the data will be available from direct sources, the investigator is ready to enter into the detailed planning of the work of collecting them.

Sometimes a combination of library sources and direct sources can be used. For example, in comparing wage rates in a particular community with rates for similar employment in the entire state and the entire country, it might be feasible to obtain the data for the state and the nation from the reports of the United States Bureau of Labor Statistics, whereas the local data would have to be secured directly from business concerns in the community. In all such cases it is desirable to make as much use as possible of library sources.

If the data required for an investigation can be obtained from library sources, the procedures discussed in the next sections will not be needed. On the other hand if direct sources must be used, the investigator must be thoroughly familiar with the principles of sampling and with the practical technique of the collection process by the method of either sample or census.

Census and Sample

In some investigations it is desirable or even necessary to make a complete enumeration. This is known as the census method. The census method is used in part of the statistical work of the federal government. The decennial population census is a complete enumeration, as are the Census of Manufactures, the Census of Business, the Census of Agriculture, and others. Other complete collections of data are by-products of the tax-collecting function of the government. Examples of these are the statistics of imports, corporate and individual incomes, cigarette consumption, and gasoline consumption.

In contrast to these cases of complete collection of data are the

great majority of external investigations in which the census method is impossible. Instead of collecting all of the information concerning a given subject, these investigations depend upon obtaining a sample which will be representative of the whole. The methods of securing a representative sample will be discussed in detail in chapter V. We are interested here merely in pointing out that results representing a large population of items can be obtained by the use of a sample. In constructing a wholesale price index no attempt is made to include the price at which every wholesale transaction is made. The prices of only a few articles in important markets are used. The American Experience Mortality Table, giving the age at death of an initial 100,000 persons at age 10, was constructed from a large sample of insured lives. Crop reports of the Department of Agriculture are based upon information received from local reporters in all parts of the country who have in the aggregate a knowledge of the condition of no more than 3 to 5 per cent of the acreage planted.

Internal investigations should use complete enumeration when feasible because of the greater accuracy, but situations may arise in which the size of the investigation or the difficulty of securing data even within a single concern preclude the use of all of the data. A good case of partial enumeration occurred when a large mail-order house desired to have a check at two weeks' intervals on its gross profit or difference between total income from sales and cost of goods. The income from sales could be obtained from the accounting department, but to get the cost of goods sold in any two weeks' period would have been impossible from the point of view of both time and expense. The concern therefore took 100 orders at random³ and computed the cost of the goods included in those orders. The results were applied to the 19,000 orders which the concern filled during a two weeks' period.⁴ The method used did not lead to an exact answer to the problem, but the results were good enough to allow a current check on selling prices. In any event the time involved in using complete data would have forced the company to abandon the idea.

³ A random sample results from taking individuals from a group by some system that is in no way dependent upon the characteristics of the items chosen, so that the presence in the sample of any particular item is left entirely to chance. In this case the equivalent of a random sample could be obtained by taking every 190th order that appeared on the sales record. A similar result could be achieved by taking the first 10 orders recorded each working day of the two weeks' period. Any methods similar to these would serve the purpose of providing a sample of 100 orders which would be representative of the 19,000 filled during the period.

⁴ Example taken from page 76 of M. A. Brumbaugh and R. Riegel, *Study Problems in Business Statistics*. New York: American Book Co., 1935.

A decision must always be made on the question of census or sample. Sometimes attendant circumstances will make the decision almost automatic; in other instances they may complicate it. A case in point is the unemployment survey made at the close of 1937 by the federal government. The necessity of having quick results made it desirable to rely on a sample. On the other hand previous experience indicated that only a complete enumeration would be reliable. A compromise plan was used in which the reporting form was distributed by mail carriers to every family, and various channels of publicity were used to urge all unemployed persons to fill out the form and return it to the local post office. There was considerable doubt whether the reporting would be complete, so a check was made by house-to-house canvass in selected cities, villages, and rural areas. The check showed that the voluntary reporting was about 72 per cent complete, but that there was considerable variation in the completeness of registration from one locality to another. The sacrifice of correct method to secure a quick report lessened confidence in the result.

The Collection Method. Agents and Mail Questionnaires

If it has been determined that a complete census must be taken, there is practically no choice as to method. Only by the use of agents for personal interviews and follow-up visits can 100 per cent collection be guaranteed. Even with the aid of compulsion by the federal government, when data for the Census of Manufactures are collected by mail, it is necessary to send agents to secure delayed reports.

When sampling is deemed satisfactory as a method of investigation, there are alternative methods of approach. In a study of limited scope, one investigator may make the plan and collect all the data personally, but as a rule some other method of collection must be employed. Agents may be sent out to secure replies to a list or schedule of questions, or the personal element may be abandoned entirely in favor of the use of questionnaires sent through the mail. The two methods are sometimes combined when mail questionnaires are sent out to all from whom information is desired and after a reasonable period of time has elapsed agents are sent to those who have not replied. A variation of this method is employed when agents collect data in thickly populated centers and mail questionnaires are sent to respondents in less accessible regions. The detailed methods of collecting information by using agents or the mail will be presented in chapter VI, but the

decision as to which method to employ is an essential part of the preliminary plan and rests upon a number of considerations.

Importance of Personal Element.—The function of the agent is to create a favorable attitude toward his mission, to explain doubtful points concerning the investigation, to encourage the informant to provide the desired information, and to record responses. These things cannot be done with a mail questionnaire. It loses immediately whatever value inheres in the personal contact between an agent and an informant. The form and tone of the mail questionnaire should be designed to supply as far as possible the missing personal element, but the fact remains that a mail questionnaire is an impersonal appeal for information and the investigator must expect it to be treated as such.

Area Covered.—Investigation within a single city or local area can usually be done more thoroughly and in less time by agents. The agents can also be directly supervised and their completed schedules checked as they are turned in. All of these things add to the accuracy of the work. When a larger area is to be covered, direct supervision is impossible, a larger number of agents cannot be as well trained, and the value of direct contact with the informant is greatly diminished because of poorer agent technique. In general when a large area is to be covered mail questionnaires should be used, whereas agents are superior for investigations confined to small areas. There are, of course, exceptions to this rule as the subsequent discussion will show.

Time Element.—It is very difficult to confine an investigation using mail questionnaires to a fixed time period. The questions may require only five minutes to answer, and the need for immediate answers may be quite clear to the respondents, but actual experience shows that the replies will straggle back over a period of time. The usual procedure is to close the collection arbitrarily after a reasonable period has elapsed and enough replies have been received to permit analysis. The more involved the questionnaire, the greater the uncertainty as to when the replies will be received. In planning the successive steps of an investigation using mail questionnaires, it is never advisable to allot a certain period such as two weeks or one month for the replies to be returned. Unless some flexibility is introduced into the time plan, the subsequent steps of the work are likely to be disorganized by unexpected delay in receiving filled-in questionnaires.

An example from the writer's experience will illustrate what can happen. An association of worsted yarn manufacturers requested an

investigation of their equipment and trends in the production of different kinds of yarn. A questionnaire was prepared and submitted to the supervising committee of the association for approval. After making several suggested changes the form was printed and distributed to the 96 members of the association during a meeting at which practically all members were present and agreed to co-operate by returning the information within 30 days. At the end of 30 days about 20 completed questionnaires had been received. Another month elapsed during which an additional 10 or 12 replies were received. A "follow-up" letter was then sent to all delinquents. This brought another 20 replies within a month. During the next six months cajolery, personal visits, and personal favors brought the total of completed replies to 70. The work of analysis was completed just about one year after the investigation was initiated.

In contrast to the uncertainty encountered in this example, the use of agents permits the establishing of a definite time schedule. The agents can be allotted fixed amounts of work and their operations can be carefully supervised. If two months, for example, have been allowed in the plan for agents to collect the data, it can be definitely expected that at the end of the two months all reports will be turned in. Nothing as certain as this can be anticipated when mail questionnaires are used.

Percentage of Replies.—Where agents are used an investigator can lay his plans to get a certain number of cases and enough agents can be put in the field to secure the desired number within a specified time. No equivalent certainty concerning the number of cases can be introduced when mail questionnaires are used. Usually a large part of those to whom questionnaires are sent will disregard them; hence a return of 10 to 20 per cent on an ordinary investigation is the likely response. However, there are in particular cases circumstances which may result in a much higher or much lower percentage of return. Actual experience with questionnaire technique leads to certain general explanations of the small proportion of replies. These can be listed as follows:

a) Some individuals and certain classes of the population have an aversion to giving any information under any circumstances. Others intend to reply but fail, due simply to inertia.

b) The questionnaire method has been over-used to such an extent that busy men throw all questionnaires into the wastebasket.

c) The sponsorship of a well-known individual or agency may increase the percentage of replies, and the absence of any such identification may affect the percentage of replies adversely.

d) As a rule, the shorter the list of questions, the higher will be the percentage of replies.

e) Simple questions with "yes" or "no" answers will bring a better response than complicated questions.

f) When the respondents have a direct interest in the subject matter of the questionnaire, or when they will receive some personal or group benefit such as a premium, a free sample, or a copy of the results of the study, the percentage of replies will be above the average.

These are some of the factors that lead to the low percentage of replies received from a mail questionnaire and they must be taken into account in estimating how many questionnaires should be sent out in order to get a desired number of replies.

Cost.—The question of cost is closely related to area covered. In a local investigation the agents can be assembled for training at nominal expense and their transportation costs while in the field are small. In a larger investigation centralized training means transportation for the agents to the training point and back to the field, while decentralized training involves transportation for the training staff or the establishing of a number of training staffs. All of this is not only expensive but extremely cumbersome.

In an investigation covering a large area, mail questionnaires are usually less expensive than schedules collected by agents. Suppose that 8,000 letters were sent out in an investigation and 1,500 replies were received. The cost except for preparation of the questionnaire would be about as follows:

8,000 envelopes	at \$4.50 per thousand	...\$ 36.00
8,000 business reply envelopes	at \$6.50 per thousand	... 52.00
8,000 addressing, folding and insertion	at \$2.25 per hundred	... 180.00
8,000 stamps	at \$.03 each	 240.00
1,500 stamps (business reply rate)	at \$.04 each	 60.00
Total cost		\$568.00
Cost per reply ($568 \div 1,500$) = \$.38.		

The estimated cost turns out to be 38 cents per questionnaire which is probably less than the cost of using agents.

On the other hand it does not always follow that agents are cheaper for a local investigation. If this same study were made within a single

city, the expense of postage would be reduced making the cost about 32 cents per questionnaire. If a corresponding estimate of the cost of using agents turned out to be more than 32 cents per schedule, then mail questionnaires would be cheaper even though the investigation were local in scope.

Amount and Complexity of Information.—If the number of questions is small, answers can be obtained by mail. A long list of questions practically precludes the use of the mail questionnaire because too few replies are likely to be received. To get replies to a long list of questions requires the persuasion of personal contact between agent and informant. Also the information which can be obtained by mail must be relatively simple. Questions which require lengthy explanations or interpretations or information which is difficult for the respondent to give, particularly if long statements are necessary to answer questions, all tend to reduce the number of replies by mail. When an investigation involves asking questions of this sort agents should be used. Replies by mail will not be satisfactory either in number or in accuracy.

A practical example will illustrate the circumstances under which one method is more suitable than the other. A research bureau collects retail prices of food articles monthly from 25 grocery stores and monthly sales from 50 drugstores. All of these stores are located in one city, yet the bureau uses agents to collect food prices but mail questionnaires to collect drugstore sales. The difference lies in the fact that any clerk can give the food prices or the agent himself can take them from the price tags. Also the food prices are available any time the agent appears at the store, whereas the sales figures for drugstores are not made up until the manager or owner has time to prepare them. An agent might have to make several visits for the data. Further, the grocer would not bother to write down the prices of the 42 articles which appear on the schedule, but the druggist does not object to transferring a single sales figure from his ledger to the bureau's collection sheet. These examples illustrate the kinds of facts which can be obtained by agents and by mail questionnaires.

Type of Information.—Quite apart from the question of complexity there are certain types of information which can be obtained better by mail, others are more suitable for collection by agents. Mail questions must not offend. The same, of course, is true of the questions on a

schedule in the hands of an agent. However, the agent can get personal information which cannot be obtained by mail. Skillful interviewing may procure confidential information on subjects which would be offensive in the absence of the personal element. In 1936 the United States Public Health Service collected data by the use of agents from thousands of families all over the country on subjects entirely beyond the reach of mail questionnaires. Here are some examples from that schedule:

1. What disabling illness occurred in the family during the past year?
2. Is there other handicapping disease or condition?
3. Has anyone in this home ever been examined for tuberculosis?
4. Has anyone in this home been to a health clinic or health center during the past year?
5. Is any member of the family crippled, deformed, or paralyzed?
6. What is the annual family income?

This schedule had 64 questions, most of them on a par with those given, which in each instance asked for details as to conditions, treatment, and physician in attendance. Information of this sort could not have been obtained by mail.

Bias.—When questionnaires are mailed to a list of business concerns or persons, that list has been selected as a representative group from which to obtain the desired information. At that point, however, the investigator's control over the group ceases. Some will reply, others will not. Are those who reply representative of the entire group? Experience shows that when a request is made to business men many who are not able to make a favorable report on the information requested will not reply at all. This tendency introduces a definite bias into the results and greatly reduces the value of the questionnaire method. An equally disconcerting bias enters when questionnaires are sent to individuals. Those with more education or more experience are likely to reply, whereas whole segments of the population which one may wish to reach will disregard the questionnaire entirely. In general then, a bias is likely to appear in the replies to a questionnaire because the ones who reply are not representative of those to whom the questions were sent. Notice that this is quite apart from any tendency of respondents to give biased answers, a difficulty which the investigator faces whether the data are collected by agents or by the use of questionnaires.

Summary.—In deciding whether to use agents or mail questionnaires in a particular case, all of these factors must be taken into account. Sometimes one will be determining, again the balancing of all of them will point to the preferable method.

Occasionally there is an advantage in using a combination of the two methods. A schedule of questions can be sent to the informant by mail with a request that it be given preliminary consideration pending the arrival of an agent at a later date. This method is effective in investigations requiring complex information or where it is necessary to assemble the information from various offices of a business concern. The work can be done prior to the arrival of the agent, but the agent can go over the schedule to be sure the questions have been interpreted correctly. A modification of this method is used by the Department of Commerce in taking the Census of Business.

PREPARE A STATEMENT OF THE PROGRAM

In the course of attending to the various details arising in the preceding steps there is a chance that some essentials may have been overlooked, or that points originally included in the plan will subsequently be forgotten. To prevent such contingencies the entire program should be put in writing. There are several advantages to the investigator in doing this. It forces him to regain proper perspective with respect to the investigation. It permits him to pick up any loose end in his plan. It gives him a complete statement to which reference can be made in the future, if puzzling situations arise. It provides a preliminary outline for writing the final report.

The statement of the program should be submitted to the sponsor⁵ of the investigation for approval. This step is particularly necessary when the project involves direct collection of data but is applicable to some extent even though the data are to be taken from library sources. All too frequently misunderstandings between investigator and sponsor arise subsequently because of failure to come to an agreement at the beginning as to exactly what will be done and how it will be done. The investigator should present his program in writing and in return insist upon a written approval from his sponsor.

⁵ "Sponsor" means the organization or individual authorizing the investigation. Hence the sponsor may be a board of directors, a board of trustees, a higher executive, an advertising agency, etc.

PROBLEMS

1. Each of the following is a statement of a problem for investigation. Rewrite any of them that fail to define the problem completely.
 - a) Retail sales taxes have no adverse effect on the sales of cigarettes.
 - b) Between 1938 and 1941 the movements of prices on the New York Stock Exchange can be explained largely by charting with them the series of crises and tensions in European affairs.
 - c) We (management) know that the change in the time of introducing new models of the Kistler automobile from January to November has changed the sales curve, but we are in doubt whether the expected decrease in the peak and trough of sales has occurred. Prepare a report on this question for the meeting of sales representatives on September 23.
2. A student was given the following assignment in a statistics class in 1935. "Has the center of the slaughtering and meat packing industry been moving westward during the past 40 years?" The student read *The Jungle* by Sinclair, a story of conditions in the industry in Chicago. He then proceeded to collect data on the number of head of cattle, sheep, and hogs shipped from each state of the United States, and the number of animals slaughtered at various important cities, such as Omaha, Kansas City, Chicago, and Buffalo. He also collected data on the livestock receipts at principal markets. The student then sought help in completing the work.
 - a) What criticism would you make of his work to date?
 - b) How would you advise him to proceed?
3. Which of the following are library and which are direct sources?
 - a) The price of wheat is obtained from a daily paper.
 - b) The sales of retail drugstores in a community are reported by the individual stores monthly to a research bureau which issues a monthly report of combined sales to the reporting stores and to newspapers.
 - c) An advertising agency calls residences by telephone to inquire whether the radio in the home is in use.
 - d) The federal income tax law requires that a copy of all tax returns be kept on file for public inspection in local offices of the Bureau of Internal Revenue. A student prepares a study of income distribution in his city based on these duplicate tax reports.
4. Investigate each of the following in the reference given to determine whether the method of collection is by sample or census.
 - a) Each year, beginning in 1935, the Department of Agriculture publishes complete information on agricultural production. *Agricultural Statistics*, United States Department of Agriculture, pp. 1-5 (approximately).
 - b) The net profits of corporations as compiled by the Federal Reserve

Bank of New York. *Survey of Current Business*, 1938 Supplement, United States Department of Commerce, pp. 64 and 180.

- c) The value of production of manufactures in the United States. *Biennial Census of Manufactures*, any issue, United States Department of Commerce. (The description of method is found at different places in different issues. In the 1925 Census, for example, the description of method is found on pp. 3-6.)
 - d) The loans and investments of reporting member banks of the Federal Reserve System in 101 cities. *Survey of Current Business*, 1938 Supplement, United States Department of Commerce, pp. 55 and 178.
5. State in each of the following examples of collection whether agents or mail questionnaires should be used and whether the census or sample method should be used. Give reasons for answers in each case.
- a) A city welfare organization wished to make an investigation of the extent to which families receiving city relief were paying money on installment purchases.
 - b) A city restaurant association wished to study the distribution of expenses of doing business of its 53 members.
 - c) An advertising agency wished to inquire from the owners of a certain make of automobile whether they would purchase the same make of car again.
 - d) A corporation wanted information concerning how many of its 4,500 employees were home owners, the value of their homes, and where the homes were located.
6. In each of the following examples the student is expected to lay out a preliminary plan for the collection of data, giving explanations of procedure and reasons for choice where alternate methods are available.
- a) A study of vacant dwellings in the community in which your college or university is located. The purpose of the study will presumably be to determine: (1) the percentage of dwellings vacant, (2) what types of dwellings have the highest and lowest vacancy ratios, (3) the sections of the community having the highest and lowest vacancy ratios, (4) the relation of vacancy to age of dwellings, (5) allied questions that you may care to include.
 - b) An automobile manufacturer advertising in newspapers and magazines, on billboards, and by radio wishes to discover which type of advertising is most effective in drawing the attention of the public to his product.
 - c) A manufacturer of a well-known brand of toilet soap wishes to discover by a direct appeal to consumers why the sales of his product have declined during the past year.
 - d) A state milk control board wishes to find the variations in the price at which whole milk is sold in retail stores in the state.

REFERENCES

BOWLEY, ARTHUR L., *Elements of Statistics*. London: P. S. King and Son, Ltd., 1920 (fourth edition).

Pages 14 and 15 contain a brief but effective statement of the preliminary planning of statistical investigations.

BROWN, LYNDON O., *Market Research and Analysis*. New York: The Ronald Press Company, 1937.

Chapter 8 gives a brief statement of the fundamentals of planning an investigation.

CHAPIN, F. STUART, *Field Work and Social Research*. New York: The Century Company, 1920.

Chapters I, II, and III are devoted entirely to the preliminary planning of investigations.

EIGELBERNER, J., *The Investigation of Business Problems*. Chicago and New York: A. W. Shaw Co., 1926.

Chapters I-VI provide general background for the principles of collecting data.

SANDERS, ALTA G., and ANDERSON, CHESTER R., *Business Reports*. New York: McGraw-Hill Book Co., Inc., 1929.

Chapters VI and VII give a detailed statement of the preliminary work involved in the collection process.

SCHLUTER, WILLIAM C., *How To Do Research Work*. New York: Prentice-Hall Inc., 1929.

Chapters I-X contain a painstaking description of the preliminary steps of investigation.

SPAHR, WALTER E., and SWENSON, RINEHART J., *Methods and Status of Scientific Research*. New York: Harper and Bros., 1930.

Chapter X contains a summary of methods of collecting data.

CHAPTER V

SAMPLING

RELATION TO KNOWLEDGE

MUCH of the world's knowledge is based upon inferences drawn from observation of samples. Finding the skeleton of a giant mammal embedded in rock strata demonstrated to have been on the surface of the earth 100,000 years ago, the paleontologist deduces the fact that such an animal lived in that period and then generalizes that this animal was typical of many alive at the time. One example has been found; therefore many others like it must have existed. A lumber jack taps along the side of a fallen tree with his axe and, listening to the sound, determines how far the tree is hollow and just where it becomes solid to the heart. His past experience in tapping logs represents a large sample providing the knowledge to be applied to the new log and his judgment is usually correct. A public speaker, wishing to drive home a point to his audience, illustrates with a story or an experience because he has found that this method of emphasis is the most effective. The response he has obtained in the past represents a sample the results of which are a part of his platform technique. These illustrations exemplify the extent to which sample experience becomes the guide to current action. Similar examples could be cited in every field of knowledge. The notion of sampling and the generalization of the results of sampling are in no sense peculiar to statistical work. Sampling is, however, particularly important in the field of statistics, because the numerical character of the subject lends itself to exact development.

THE IMPORTANCE OF SAMPLING

Sampling techniques are seldom necessary in internal statistical work but have their greatest application in external work. In the latter case it is seldom possible to obtain all of the data pertinent to a given statistical universe,¹ hence the usual situation requires that results be

¹ The complete category of data from which a sample is drawn is known as a statistical universe or statistical population. In the preceding chapter an example was presented in which prices of groceries were collected monthly from 25 grocery stores in a city. The 25 stores are a sample of the universe or population consisting of all grocery stores in the city.

obtained from the study of samples. Thus we have: indexes of commodity prices based on a few hundred of the thousands of commodities that are traded daily; average hourly wage rates in manufacturing plants determined from samples including no more than 5 per cent of factory workers; the market for a product estimated from the results of sending a questionnaire to 1 or 2 per cent of the potential users. These examples are sufficient to indicate the importance of sampling in statistical work.

When a retailer buys shoes from a salesman he expects that they will be just like the sample which the salesman shows. In the same way we might expect to estimate the average age of 2,000 freshmen in a university from the ages of a sample of 100 of them attending a freshman lecture. Certain differences between these two "samples" will immediately occur to the reader. The shoes are all made on the same machinery, defects are weeded out by inspection, and uniformity is assured at every step of the manufacturing process. Therefore, one shoe picked at random does represent the entire lot. On the other hand we cannot be sure that the sample of 100 freshmen is representative as to age. The lecture may have attracted only more mature students, or a brilliant younger group who completed high school in three years.

The absence of control over statistical data is precisely what makes it necessary to develop principles and methods of sampling. We wish to know something about a certain universe of events or facts, but are unable to make a complete enumeration. Instead we must record specific facts concerning a sample drawn from the universe—a sample which shall be representative of the universe. The problem is, How can such a sample be obtained?

THE PRINCIPLE OF STATISTICAL REGULARITY

If our knowledge of the universe is limited how can we ever know that a sample drawn from it is representative? The answer to this question comes from a principle which is as broad in its application as the laws of nature. It is known as the Principle of Statistical Regularity and may be stated thus: *A sample selected at random from a universe will exhibit the characteristics of the universe, even though the number in the sample is small compared with the universe.* The simplest illustrations of the operation of the principle occur in coin

tossing and dice rolling. Every throw is exactly like every other one and the experimental material, i.e., coins or dice, remains constant.

The result of an experiment with coins is presented in Table 6. Ten coins were used and the results in the first four lines of the table are for groups of 50 throws each, or 500 coins. The 240 heads and 260 tails obtained in the first trial of 500 varied 4 per cent from the expectation of 250 each. The second trial gave 245 heads and 255 tails, a cumulative result of 485 heads and 515 tails in the first 1,000 coins. The cumulative result varies 3 per cent from the expected 500 of each. In successive rows of the table the results for the third and fourth trial of 50 throws, the third and fourth hundred throws, and

TABLE 6

THE PRINCIPLE OF STATISTICAL REGULARITY ILLUSTRATED BY COIN THROWING

NUMBER OF THROWS OF TEN COINS EACH	RESULT		CUMULATIVE			
			Actual Result		Expected Result (Equal Number of Heads and Tails)	Per Cent Variation from Expected Result
	Heads	Tails	Heads	Tails		
1st 50	240	260	240	260	250	4.00
2d 50	245	255	485	515	500	3.00
3d 50	253	247	738	762	750	1.60
4th 50	246	254	984	1,016	1,000	1.60
3d 100	501	499	1,485	1,515	1,500	1.00
4th 100	539	461	2,024	1,976	2,000	1.20
1st 400	2,024	1,976	2,024	1,976	2,000	1.20
2d 400	1,925	2,075	3,949	4,051	4,000	1.28
3d 400	1,999	2,001	5,948	6,052	6,000	.87
4th 400	2,036	1,964	7,984	8,016	8,000	.20
5th 400	2,009	1,991	9,993	10,007	10,000	.07
6th 400	2,007	1,993	12,000	12,000	12,000	.00
7th 400	2,015	1,985	14,015	13,985	14,000	.11
8th 400	2,000	2,000	16,015	15,985	16,000	.09
9th 400	1,993	2,007	18,008	17,992	18,000	.04
10th 400	1,959	2,041	19,967	20,033	20,000	.17
11th 400	2,013	1,987	21,980	22,020	22,000	.09

so on are shown. The last column of the table shows how the percentage variation from the expected number tends to decrease as the size of the cumulative sample increases. At the end of the first 400 throws (4,000 coins) the variation is 1.20 per cent. At the end of the second 400 throws the variation increases slightly to 1.28 per cent and then decreases regularly through the third, fourth, and fifth groups of 400 throws and reaches zero at the end of 2,400 throws. The exact result obtained at this point is purely accidental, as is the exact way in which the percentage variation declined with the increase in size

of the sample. The important point is that the percentage variation becomes smaller and smaller as the size of the sample increases and that in spite of slight deviations it remains small through the seventh, eighth, ninth, tenth, and eleventh trials of 400 throws each.²

Examination of the result columns shows that sometimes the number of heads is greater than the expected number and at other points the number of tails is greater than expected; there is no indication of any fixed bias nor any tendency for either heads or tails always to exceed expectancy. Specifically, for instance, after 400 throws there were more heads than expected, but after 800 throws there were more tails than expected. This difference in the direction of variation together with the regular reduction in the percentage variation indicates the tendency toward regularity of the results as the size of the sample increases. The tossing of 44,000 coins has demonstrated the principle. If the tossing were continued, the percentage variations could be expected to diminish.

A further demonstration of the operation of the principle of statistical regularity is presented in Table 7, showing the results of throwing five dice. The first line of the table gives the results of the first 20 throws of five dice (100 faces showing or, as recorded in the first column, 100 dice). The maximum variation from the expected number is the appearance of 22 fives. This variation of 32 per cent is due to the small size of the sample. The decline in the variation of the actual from the expected result can be seen as the size of the sample is increased.

There are two places at which the progressive decline in variability is broken—when the cumulative sample consists of 300 dice and when it consists of 3,600 dice. These two exceptions to the operation of the principle of statistical regularity do not disprove its universality. The experiment was carried on with ordinary commercial dice and they were thrown within a confined space rather than being permitted to come to rest without obstruction. Either circumstance might be sufficient to explain the two irregularities that appear.

In both examples the expected occurrence of the recorded events is known. That is, coins should fall heads and tails with equal frequency; one face of a die is as likely to turn up as another. Consequently

² The reliability of a sample increases proportionally to the square root of the number of cases in the sample. Thus to double the reliability, i.e., to halve the variability, the number of cases must be four times as great. The reason for this relation will be apparent from the form of the formulas for standard error in chapter XXIX.

TABLE 7
THE PRINCIPLE OF STATISTICAL REGULARITY ILLUSTRATED BY DICE THROWING

No. of Dice Thrown	Expected No. of Times Each Face Should Appear	No. of Times Each Face From 1 to 6 Appeared						Cumulative No. of Times Each Face From 1 to 6 Appeared						Expected No. of Times Each Face Should Appear	Maximum Per Cent Variation From Expected
		1	2	3	4	5	6	1	2	3	4	5	6		
1st 100.....	16 $\frac{2}{3}$	14	18	18	15	22	13	14	18	18	15	22	13	16 $\frac{2}{3}$	32.00
2d 100.....	16 $\frac{2}{3}$	21	20	14	14	18	13	35	38	32	29	40	26	33 $\frac{1}{3}$	22.00
3d 100.....	16 $\frac{2}{3}$	18	16	15	23	17	11	53	54	47	52	57	37	50	26.00
2d 300.....	50	57	51	51	46	50	45	110	105	98	98	107	82	100	18.00
2d 600.....	100	92	111	109	92	92	104	202	216	207	190	199	186	200	8.00
1st 1,200.....	200	202	216	207	190	199	186	202	216	207	190	199	186	200	8.00
2d 1,200.....	200	193	197	189	207	204	210	395	413	396	397	403	396	400	3.25
3d 1,200.....	200	178	188	216	220	208	190	573	601	612	617	611	586	600	4.50
4th 1,200.....	200	214	184	213	203	186	200	787	785	825	820	797	786	800	3.13
5th 1,200.....	200	213	196	181	208	207	195	1000	981	1006	1028	1004	981	1000	2.80
6th 1,200.....	200	192	199	204	196	201	208	1192	1180	1210	1224	1205	1189	1200	2.00
7th 1,200.....	200	207	207	205	185	182	214	1399	1387	1415	1409	1387	1403	1400	1.07

these are controlled experiments carried out to show how the principle of statistical regularity operates.

Consider another example. A teacher made a practice each year of having each of his students measure the width of his desk. The same ruler was used year after year, but the students' results varied individually and from year to year. The ruler was subdivided by 32ds of an inch and the students were told to read to 64ths of an inch. The yearly average for eight years is shown in Table 8.

TABLE 8
WIDTH OF A TEACHER'S DESK ACCORDING TO MEASUREMENTS BY
EIGHT DIFFERENT GROUPS OF STUDENTS

YEAR	NUMBER OF STUDENTS	AVERAGE MEASURED WIDTH OF DESK IN INCHES
1st	32	48.65
2d	26	48.49
3d	30	48.61
4th	31	48.60
5th	28	48.57
6th	36	48.59
7th	31	48.62
8th	28	48.60

The exact width of the teacher's desk is unknown, yet these averages perform in the same way as the observations of coin and dice throwing. As the number of heads sometimes exceeded the number of tails and vice versa, so in the same way some of these averages were slightly above the theoretically true width,³ others slightly below.

The example shows that results from samples tend to group themselves about an unknown true value just the same as they group themselves about a known true value. The coins and dice are samples in which the observations involved only counting. The desk data involved measurement. We conclude therefore that the principle of statistical regularity applies to both counted and measured samples.

A major question arises at this point. Will the same regularity appear when the universe is less homogeneous⁴ than coins, dice, or

³ The theoretically best estimate for an unknown value is the arithmetic average of a large number of independent observations of that value.

⁴ "Homogeneous" as used in statistical work means sufficiently alike to be used for the immediate purpose as though equivalent. For example, coins and dice are truly homogeneous in the sense that each one is identical with every other one, but human beings, animals, and other materials dealt with in statistical work are treated as though they were homogeneous even though appreciable differences of size, weight, and other characteristics appear within the groups. "Non-homogeneous," or "heterogeneous," means possessing characteristics which are sufficiently different to require classification in different categories.

the width of a desk? The answer can be obtained from another example. The problem was to find the average number of letters in the last names of persons having telephones in Buffalo, New York. Each page of the telephone book contained four columns. A ruler was laid across each page of the book near the middle of the page and the number of letters in the name appearing above the ruler in each column was counted. Four samples were taken, the first containing a name from the first column of each page of the book, the second a name from the second column and so on. There were 265 pages in the book, hence each sample contained 265 items. The average numbers of letters in the names in the samples were:

1st sample	6.51 letters
2d sample	6.52 letters
3d sample	6.51 letters
4th sample	6.54 letters

The similarity of the four results shows how the principle of statistical regularity operates. These samples were chosen entirely at random,⁵ yet any one of them alone presumably would have represented the universe.

Each sample contained only 265 out of a total of about 70,000 names in the telephone book, yet there is little doubt that the average number of letters in the last names appearing in the book is about 6.5. Note that we do not expect the sample to give the exact characteristics of the universe but rather an approximate indication of those characteristics. The four samples vary slightly and probably each one varies somewhat from the true value. Such variations will always appear in samples. In fact the methods of analyzing data which will be developed in subsequent chapters include the measurement of the expected variation of the characteristics of a universe from those characteristics found in a sample drawn from it.

In the previous examples chance has operated in each case. The chances are equal of getting either heads or tails in tossing a coin; the chance that any one face of a die will turn up is one-sixth. In measuring the width of a desk overestimates and underestimates are equally likely. In the telephone book there is just as much chance that "Fry" will be printed near the middle of the page, as "Freundenberger."

| The cases which arise in practical business affairs are usually not

⁵ The concept of random selection of cases for a sample is explained in chapter IV, p. 58, footnote 3.

so simple as these examples. The operation of pure chance which is so evident in the examples will be for the most part lacking in practical work. The investigator is forced to deal with conditions as they exist. In general more variables will be present and as a result adjustments become necessary. The real problem of sampling is to find methods of selecting the cases for the sample so that the characteristic to be measured or counted has a chance of occurring in the sample in the same proportion as it occurs in the universe. The amount of care required to do this will be evident when it is remembered that the extent of occurrence of the characteristic being studied is unknown in the universe and can only be inferred as the final step in the analysis of the sample. This would be circular reasoning were it not for the principle of statistical regularity. Some of the characteristics of the universe may already be known, and if these *known* conditions of the universe can be reproduced on a small scale in the sample, then the operation of the principle is all that is needed to allow us to infer from its occurrence in the sample the extent to which a given unknown characteristic is present in the universe.¹

THE TWO PROBLEMS OF SAMPLING

There are two major factors to be considered in obtaining a sample:

- (1) how many cases must be included to obtain reliable results and
- (2) what cases must be included to secure representativeness.

The Size of A Sample

The first problem is, How many or what proportion of the cases in the universe must be taken for the principle of statistical regularity to operate? There is no numerical answer to this question. It would be wrong to say that a 50 per cent sample or a 10 per cent sample will be satisfactory. In fact such an answer is meaningless in coin tossing where the universe is infinite. Even when the universe is limited, as in the telephone book example, we do not attempt to say that a certain number of cases or a certain percentage of the total number of cases in the universe will be a large number. The telephone book used in this test experiment contained about 70,000 names. A sample of 265 was therefore only about four-tenths of 1 per cent, yet the results obtained from the four independent samples fell within a very narrow range.

The question of how many cases to include in a sample must be decided for each problem separately. The number depends primarily on the degree of reliability required and the diversity of the characteristics present in the universe. The tests for reliability are developed in detail in chapter XXIX. The question of diversity of characteristics can be discussed at this point. If the universe is as strictly homogeneous as the letters in names in a telephone book, a very small sample will suffice for the purpose of determining the average number of letters per name. On the other hand, if a sample from the telephone book were used to determine the percentage of subscribers who used four-party service, a much larger sample would be required to insure that proper provision was made for the tendency toward use of this type of service in different parts of the city, for the inclusion of mainly residential subscribers since few business places use four-party service, and for the exclusion of those exchanges, if any, which do not provide four-party service.

This example demonstrates further the importance of a previous statement, that the question of homogeneity of the universe depends upon the purpose for which the sample is taken. Thus only a relative statement can be made concerning the size of a sample. If the events in the universe differ only with respect to the characteristic which is tested by the sample, a sample as small as one-tenth of 1 per cent of the universe may be adequate for the principle of statistical regularity to be effective. As the number of characteristics which vary in the universe increases, the size of the sample must be increased, sometimes becoming as large as 10 per cent of the universe. If a sample greater than 10 per cent is required to reproduce the characteristics of the universe, the universe itself is probably not sufficiently homogeneous for the principles of sampling to be used.

Methods of Securing a Representative Sample

The second problem is how to secure a representative sample. In particular, What cases shall be included in order to set up in the sample a pattern which will reproduce on a smaller scale the conditions of the universe? In some cases none of the conditions of the universe may be known, while in other instances information is already available concerning the distribution of certain characteristics. Consequently there are two methods of securing representativeness: (1) uncontrolled sampling and (2) controlled sampling.

Uncontrolled sampling.—If little or nothing is known about the distribution of any of the characteristics in the universe, an uncontrolled sample is the only one which can be used.

Example 1: Suppose a tobacco retailer wishes to make a consumer investigation of the question, "What brand of cigarettes is most popular in this city?" It would be difficult to find a "control" in this case, because nothing is known regarding the characteristics of cigarette smokers as a group of the population. It is known in general that children do not smoke cigarettes, but just what the proportion of cigarette smokers is in each adult age group would be hard to estimate. It is not even known how the percentage of men cigarette smokers compares with the percentage of women smokers. If there had been some recent nation-wide study showing what percentage of each sex smokes cigarettes, these two percentages would provide a control to determine the proportionate number of men and women from whom replies should be obtained in this study.

In the absence of any such control, the most obvious method is to take cases from the universe as they come to hand, making no choice of any kind. Even this involves some selection of time and place for taking interviews. The method must insure a rough representation of all the general characteristics of the adult population, on the assumption that, provided the sample is large enough, smokers of various brands of cigarettes will be included in correct proportion. That is, since age, sex, nationality, economic class, or other characteristics may be determining factors in the choice of brand of cigarettes, the answers must come from persons who are representative of the total adult population in as many respects as possible. It would not serve the purpose, therefore, to distribute questionnaires only at women's clubs, or only at an industrial plant, or to interview only relief clients, or only people on the street at three o'clock in the afternoon. But if a busy downtown corner were selected, at an hour when all classes of men and women, employed as well as unemployed are likely to be on the street, the passers-by should be fairly representative of the entire adult population. If stopped and asked what brand of cigarette they buy, some would reply, others would ignore the question; some would be cigarette smokers, others would not smoke cigarettes or would not smoke at all. The investigation might show that six hundred and thirteen cigarette smokers replied and the answers were tabulated as follows: Brand "A," 18 per cent; Brand "B," 16 per cent; and so on.

¹ The important feature of this method is the absence of control of the sample. Experiments have shown that reliable results can be obtained by this method only through the use of a large sample. The distribution of none of the characteristics in the universe is known; therefore a large sample must be taken to insure that the pattern of the universe will be reproduced. This uncontrolled plan of collection is known also as the *extensive method of sampling*.

Example 2: The use of this method in an investigation is illustrated by the chain store inquiry of the Federal Trade Commission conducted in 1928 and published in 1930-31. The sample was obtained in the following manner:

A mailing list of chain stores was prepared by the commission from various lists of chains, including those of the Chain Store Age and the National Association of Real Estate Boards, supplemented by telephone directories, trade journals, and city directories, all of which were checked to eliminate, so far as possible, duplications. When completed, this mailing list for the selected groups of chain stores included slightly over 7,500 names. The results obtained from this mailing list are shown in the following tabular statement:

Schedules mailed	7,515
Returned by post office.....	713
Duplications	638
Non-chain establishments only.....	1,282
Co-operative group only.....	39
Reported out of business.....	492
In receivership, no records, or records destroyed, etc.....	833
Unobtainable at time of tabulation.....	1,596
Total eliminated	<u>5,593</u>
Schedules returned	1,922 ⁶

Only 1,727 of the 1,922 schedules were usable in the analysis but the Commission appraised their representativeness as follows:

Comparing the commission's data with estimates for the entire field based upon census data, it appears that the commission's study represents approximately one-half of the number of stores operated and one-half of the aggregate sales volume of all organizations engaged in chain-store merchandising in 1929 in the 26 kinds of business covered by this inquiry, including chains of two and three stores, which are not classed by the census as chain stores. On the other hand, the total number of chains represented in the commission's inquiry is estimated to be something under 10 per cent of the total.⁷

⁶ "Scope of the Chain-Store Inquiry," *Chain Stores*, 72d Congress, 1st Session, Senate Document No. 31, p. 9.

⁷ *Ibid.*, p. ix.

The Commission had to treat its data in different classifications, hence the real problem of representativeness arose in the sub-groups. A comparison of sample data with Census of Distribution data based on the parts of each which were considered comparable is shown in Table 9.

TABLE 9

PERCENTAGE OF TOTAL CENSUS CHAINS (FOUR STORES AND UP), STORES, AND SALES IN 1929 REPRESENTED IN THE COMMISSION'S ORIGINAL CHAIN-STORE SCHEDULE RETURNS FOR CHAINS OF SIX STORES AND UP FOR 1928*

KIND OF CHAIN	PERCENT REPRESENTATION OF CENSUS IN COMMISSION'S SAMPLE		
	Chains	Stores	Sales
Food	23.8	76.9	76.5
Drug	17.7	42.8	55.5
Tobacco	17.2	109.8	104.0
Variety	44.9	80.4	89.8
Clothing, furnishing, and accessories.....	20.8	30.6	34.7
Hats, caps, and millinery.....	20.6	33.7	52.9
Shoe	34.4	55.6	56.8
Department store and dry goods.....	28.7	53.5	97.8
General merchandise	6.4	9.3	7.8
Furniture	4.0	7.6	11.9
Musical instruments	12.1	15.7	24.8
Hardware	13.3	25.3	22.0
Total	21.8	66.3	69.2

* "Scope of the Chain-Store Inquiry," *Chain Stores*, 72d Congress, 1st Session, Senate Document No. 31, p. 28.

Table 9 is complicated by the fact that the ratios in the three columns have in the numerator results from the Federal Trade Commission's sample of chains operating six or more stores in 1928 and in the denominator results from the Census of Distribution, a complete enumeration of chains operating four or more stores in 1929. From this comparison the Commission concluded:

The purpose of Table [9] obviously is not to present an exact measure of the proportions of the commission's data either of the chain-store field as a whole or by specific commodities but rather to afford a general impression as to the kinds of business in which the commission data may be regarded as sufficiently comprehensive as contrasted perhaps with those for which the figures should be regarded merely as indicative because of the comparatively small representation in comparison with the census totals.

It should, of course, be recognized that the foregoing proportionate comparisons are approximations, both because of the variations in classification and the necessary treatment of the 4- and 5-store chains in the commission data.

In general, it appears that with the possible exception of general stores and furniture chains, the commission's reports are sufficiently adequate to provide a

satisfactory indication of chain store operations in the several kinds of business considered.⁸

This sample was obtained without exercising any control over the cases which should be included. As a result only part of the organizations to whom the questions were sent proved to come within the definition of chain merchandisers. The ultimate size of the sample was unknown until the returns had been edited. Even when the size of the sample as a whole was known the representativeness of the sample with respect to different kinds of chains was in doubt until a partial comparison could be made when the results of the 1929 Census of Distribution were published. Finally it turned out that in so far as the comparisons with census results were valid, the various lines of trade were not equally well represented in the sample, although with two exceptions the sample was considered large enough to provide representative information concerning chain-store merchandising in different lines of trade.

Controlled sampling.—When knowledge of some of the characteristics of the universe can be obtained, the usual practice is to take a controlled sample. *A controlled sample is one in which representativeness is obtained by conscious adjustment of the sample to conform to the conditions existing in the universe according to one or more known characteristics.* The *known* characteristics are not the ones that are being studied in the sample investigation. For instance, in a survey of buying habits of students at a certain university, the number registered in each class is a matter of record at the registrar's office. This known distribution can be used in selecting a representative number of sample cases from each class, and in order to check with this control each student interviewed must be asked what class he or she is in. However, the object of the investigation is not to determine the distribution of students by class, but rather to assemble, by means of sampling, a variety of hitherto unrecorded information regarding the buying habits of all the students.

The advantages of the controlled method lie (1) in the substitution of a known representativeness for one hoped for on the basis of size of sample alone and (2) in the small size of sample which can be used. If the information had been available for the chain-store inquiry to have followed this plan of sampling, the first step would have been

⁸ *Ibid.*, pp. 28, 30.

a study of existing information concerning chain stores to obtain for each line of trade covered by the investigation the best estimate of the number of chains, the proportion of large and small chains, the distribution of sales by lines of trade, and the total sales. Guided by this information the sample could have been planned so as to get the proper representation of large and small chains and of the several lines of trade. Thus a reasonable amount of control over the sample would have been exercised.

Controlled sampling may be further differentiated according to the degree of selection used in establishing the controls. If there is 100 per cent selection, leaving no elements to chance in determining the actual cases that appear in the sample, the method is called *selective sampling*. If a certain degree of selection is used, but the final determination of actual cases is left to chance, this is called the *inclusive method*. Examples of the latter will therefore cover the entire range between uncontrolled, or extensive, sampling in which the distribution of none of the characteristics of the universe is known, and selective sampling, in which each individual case is picked because of the known representativeness of its general characteristics.)

The selective method: An example⁹ will show the method by which the investigator, on the basis of his knowledge of the universe, hand-picks a small number of cases which he believes will be a representative sample.

Five years ago Mr. William Groom of the Thompson-Koch Company was interested in the possibility of measuring directly in terms of sales the effectiveness of the advertising produced by his agency. Mr. Groom selected four middle-western cities of 35,000 to 50,000 population. He planned to run his experimental campaigns in the newspapers in these cities and to measure results in terms of the sales made through the local drug stores. To this end Mr. Groom enrolled from three to a dozen or more drug stores in each of his cities and paid them to submit each month a statement of the sales made during the month of each of a number of different drug items.

In order to be able to generalize from the results, it is, of course, desirable to make any study of retail sales in communities which are more or less representative. Mr. Groom originally chose his four test towns on the basis of a personal knowledge of the communities, and a belief that these communities were fairly representative of a great part of the country.

The thought in using cities of this size and character was that they presented a mixture of urban and rural people and problems and, therefore, were more

⁹ Lyman Chalkley, Jr., "The Flow of Sales through Retail Drug Stores—A Factual Study," *Harvard Business Review*, Vol. XII, No. 4 (July, 1934), pp. 427-29.

representative of the whole country than either the larger metropolitan centers or the purely rural districts. Each city has some manufacturing, some farming, and some general business and professional activities proportioned roughly like those in the country as a whole.

Although only four cities were included in the study, the investigators expected their results to be representative of the country as a whole. Further the records of 23 drugstores out of a total of about 100 in the four cities were used. These 23 were personally selected by Mr. Groom. Finally the sales of 12 selected items became the data of the study. The *selection*, in every respect, of the cases to be included in the study is the important point of the example.

This description of the procedure immediately gives rise to three questions: (1) Were these four towns representative of the country as a whole as regards the relation of sales to advertising? (2) Were the sales of the 23 stores representative of sales of all drugstores in the four communities? (3) Were the 12 items the proper ones to study? The equivalent of these questions must be raised with respect to the plan for any selective sample. When the investigator feels that he has sufficient knowledge of his universe and of his sample to be able to answer such questions, there is some justification for the use of the selective method of sampling. Under the usual practical conditions no such assurance is possible, hence the method should be used sparingly. The danger lies in getting a biased result if the selection should go astray at any stage of planning the sample.

! *The inclusive method:* Two examples will illustrate various degrees of control in securing an inclusive controlled sample.

Example 1: An advertising firm was asked to make a quick survey of the nation-wide popularity of a certain brand of scouring powder. The firm first selected a few cities which were believed to be representative of the entire country. Local supervisors were appointed, each of whom was familiar with conditions in her own city, and they were asked to select sample areas that would be representative of nationality groups, old and new residential districts, etc., but each containing a variety of income levels. One agent was assigned to each area and was given freedom to select the housewives whom she would interview, except that her total number of interviews must be divided approximately into 20, 40, 30, and 10 per cent of four roughly defined economic classes. (In certain areas adjustments of the required proportions were made according to the known economic levels in the neighborhood.)

If toward the end of an assignment an agent found that she had secured a markedly unbalanced proportion of interviews according to economic class, she then had to exercise some degree of selection in choosing the blocks and houses to visit so that her remaining schedules would make up the deficiency. For the most part, however, if the agents used some system of random selection such as calling at every third house or canvassing every other block, they found that they usually had the right proportion of interviews without making any conscious adjustment.

This method permitted some degree of selection according to certain general characteristics within each of three controls of the investigation, (city, area, and economic group). In spite of this fact, most of the housewives interviewed were chosen solely by chance: they happened to be at home; they happened to live in a block near the car line where the agent started to canvass, etc., all of which were factors that had no effect whatever on their choice of scouring powder. The key idea in this method was that all of the types of families were given a chance to appear in the sample in proportion to the number of each type existing in the community. All of the characteristics of the universe were given a fair chance to be *included* in the sample. Beyond that point no selection was made of the individual cases that were actually taken.

Example 2: In this investigation¹⁰ by inclusive sampling selection was exercised only at the first level of the plan, and at later stages the returns were determined wholly by chance.

An investigation was made by the Bureau of Business Research [of the University of Pittsburgh] in the spring of 1931 to determine the cost and the quality of housing accommodations secured by salaried workers employed in downtown Pittsburgh. The housing status of 1,415 families was analyzed.

The data for this study were secured by means of questionnaires distributed to salaried workers through their employers. The co-operation of the following types of concerns with offices in downtown Pittsburgh was secured for the distribution of the questionnaires: two public utilities, four department stores, five financial institutions, five industrial concerns, one railroad, and two insurance agencies.

It is believed that the employees of these concerns represent a fair cross-section of the salaried workers employed in downtown Pittsburgh.

¹⁰ Theodore A. Veenstra, "Housing Status of Salaried Workers Employed in Pittsburgh," *University of Pittsburgh Bulletin*, Vol. XXVIII (June 10, 1932), pp. 1-4.

Questionnaires were distributed among salesmen, accountants, clerks, statisticians, engineers, and junior executives. In order to get replies from the type of worker selected for the study, co-operating concerns were asked to distribute the questionnaires to employees described as follows: "Heads of families engaged in clerical and executive work in downtown Pittsburgh with salaries of \$5,000 annually or less." In general the persons reporting were heads of families engaged in the designated types of work. Salaries in a number of cases were in excess of \$5,000; but such cases, if otherwise acceptable, were included in the study.

A large proportion of the 1,385 persons reporting occupations were engaged in clerical work. Those having executive, technical, selling, and accounting positions were next in order in numbers reporting. Other groups were only sparsely represented.

The universe from which this sample was taken included only salaried workers in offices in downtown Pittsburgh who were heads of families. The \$5,000 limit automatically excluded high-salaried executives. Thus the limits of the investigation were rather closely defined. The 19 firms were selected because they represented the known distribution of different lines of business in which the desired kinds of workers were employed. It was assumed that the employees of these concerns were representative of the universe as to types of work and salary distribution. Undoubtedly many failed to reply, but, since the original group was so carefully selected, the reply of any employee was just as acceptable as that of any other. As long as a sufficient number of replies was secured the information regarding housing could be considered as representing the entire group.

The weighted method: The weighted method of controlled sampling is, in its initial steps, the same as inclusive sampling. That is, a definite effort is made to secure cases in the sample that will represent the known characteristics of the universe. As a further step, however, the sample is again consciously adjusted after it has been collected in order to bring it into closer conformity with these known general characteristics. In this process none of the cases is dropped from the sample, but all are grouped and weighted in order to give each group the importance that previous knowledge of the universe indicates it should have.

This method was used in predicting the results of the presidential election in 1936, a detailed description of which is provided in an explanation of the work of the American Institute of Public Opinion (Gallup Polls).

The *weighted-sample* technique assumes that it is possible to isolate and measure factors, or groupings, which determine the distribution of the variable in question. The method of the weighted sample tries to choose from the many possibilities the important determinants of voting behavior. The sample is then constructed by preserving in the miniature population the ratios of the selected groupings which hold for the total population.

The problem of the selection of the significant groupings in the voting population was solved in this poll by experimentation. Gallup tried distributing straw-vote returns according to various factors. Those which showed an even distribution of ballots between the major candidates were discarded. Five controls were finally chosen. First, ballots returned from each state were to represent the correct proportion of the state's population to the national population. Second, the ratio of farm and city votes in each state was to be maintained. Third, the correct percentage of voters in each income group had to be represented. Fourth, the ballots returned were to reflect accurately the proportion of young people who had come of voting age since the last election. Fifth, the return was to come from the correct percentage of people who voted for Roosevelt, Hoover, Thomas and others in 1932.

The distribution of ballots in the proper proportion is, however, only half the story. Ballots leave the polling office in the proper ratio according to the factors mentioned. But the correct ratio is seldom maintained after the round trip from office to voter to office. As a rule less than one-fifth of the mailed ballots are returned and these tend to come from selected groups. People with intense opinions (reformers, arch-conservatives, radicals) are more likely to return ballots than those who are luke-warm or undecided; more highly educated and economically secure persons take a greater interest in the ballots and feel more free to answer them. The American Institute found that the largest response (about 40 per cent) came from people listed in *Who's Who*. Eighteen per cent of the people in telephone lists, 15 per cent of the registered voters in poor areas, and 11 per cent of people on relief returned their ballots. Men are more likely to reply than women.

These peculiarities in the mail response of the sampled population are counteracted in two ways: using interviewers, and adjusting the final number of ballots according to the original quota-controls. The Institute had some 200 interviewers scattered throughout the nation. The answers they gathered constituted one-third of the final return for the Institute poll. Interviewers can be used advantageously where the mail ballot is not likely to succeed: in relief districts, farms, and working class areas.¹¹

A partial description of the method by which the returns were adjusted according to the control groups may serve as a guide to the general application of this method.

The criteria used as controls are established from known data such

¹¹ Daniel Katz and Hadley Cantril, "Public Opinion Polls," *Sociometry*, Vol. I (1937), pp. 159-60.

as United States population by states, farm and non-farm population of each state, age groups of the population, and the 1932 election returns by states. The proportionate distribution of replies received is compared with the distribution in the "control" group for each of these criteria successively in order to secure for the votes cast in the straw ballot a redistribution that will be in every essential truly representative of the total voting population.

The adjustments according to four of the controls are made by states. As an example of the method of procedure, suppose that after the special interviews the straw ballots received from New York State were distributed as in Table 10, according to farm and non-farm voters.

TABLE 10
STRAW BALLOTS IN NEW YORK STATE, 1936
ORIGINAL RETURNS FROM FARM AND NON-FARM VOTERS
(HYPOTHETICAL DATA)

VOTERS	CHOICE OF CANDIDATE			TOTAL
	Roosevelt	Landon	Thomas	
Farm				
Number	2,500	1,975	25	4,500
Percentage distribution	55.6	43.9	0.5	100.0
Non-Farm				
Number	75,000	25,000	2,000	102,000
Percentage distribution	73.5	24.5	2.0	100.0
Total number	77,500	26,975	2,025	106,500

The first question is whether the proportions of farm and non-farm voters conform to the census distribution. Of the total 106,500 straw ballots cast, 102,000, or 96 per cent, were by non-farm voters. According to the census, however, the population of New York State is 94 per cent non-farm and 6 per cent farm. If the 106,500 straw ballots had been divided in that proportion, there would have been 6,390 farm instead of 4,500, and 100,110 non-farm instead of 102,000. These corrected figures are therefore substituted for the total ballots cast. They are then redistributed as to choice of candidate according to the same percentage distribution that was found from the actual ballots. That is, the percentage distributions shown in Table 10 are applied to the new totals for farm and non-farm giving the choices for each candidate by the farm and non-farm voters as shown in Table 11. The corrected state totals for each candidate are obtained by adding the corrected farm and non-farm votes in each case. It will be noted that the grand total for the state, 106,500, has not been altered.

TABLE 11
STRAW BALLOTS IN NEW YORK STATE, 1936
CORRECTED FOR FARM AND NON-FARM POPULATION;
DATA FROM TABLE 10

VOTERS	CHOICE OF CANDIDATE			TOTAL
	Roosevelt	Landon	Thomas	
Farm				
Number	3,553	2,805	32	6,390
Percentage distribution	55.6	43.9	0.5	100.0
Non-Farm				
Number	73,581	24,527	2,002	100,110
Percentage distribution	73.5	24.5	2.0	100.0
Total number	77,134	27,332	2,034	106,500

The same process can now be repeated for the other three controls within each state starting with the original returns for each of the three. The four totals thus arrived at according to the four criteria can then be averaged to give the true representation for New York State. After similar adjustments have been made for each state the results for the 48 states are ready to be combined according to the fifth criterion, the proportion of each state's population to the United States total. In this last step the number of ballots cast in each state is adjusted but the actual total number of ballots cast in the United States remains unchanged.

Through the introduction of these five independent controls, the internal distribution of cases in the sample might be considerably altered, but such alteration is made in order to adjust discrepancies between the collected information and the known occurrence of the five control characteristics in the universe. As a result of these alterations the cases are distributed in the sample so that the unknown variable characteristic, choice of candidate in 1936, can be studied without distortion arising from failure of one or more of the known characteristics to be represented properly.

Summary.—Before closing this discussion it should be pointed out that the three methods of obtaining a representative sample depend upon the principle of statistical regularity in different ways. (1) In an extensive sample the principle has its purest application when the appearance of the characteristics of the universe in the sample is left entirely to chance. (2) In a selective sample the investigator's knowledge of certain characteristics of both universe and sample is substituted for random choices. He is still depending on the sample to give

him information regarding certain *unknown* characteristics of that universe. (3) In an inclusive sample, either unweighted or weighted, the known characteristics of the universe are definitely projected into the conditions of the sample, but the appearance of unknown or uncontrolled characteristics is left to chance through random selection of the actual cases. None of these methods can be used in automatic fashion. Careful planning by the investigator will always be needed. His two most valuable assets will be experience and the exercise of good judgment.

PROBLEMS

1.
 - a) Would the members of your statistics class be a representative sample of the students of your school as to height? weight? age? grades? hair color? eye color? Discuss.
 - b) Would the members of the class be a representative sample of all college students with respect to the characteristics listed? Discuss.
2. If a large number of samples, each including 400 cases, show an average variability of 4 per cent from a known result, how large a sample would be required to confine the variability to 1 per cent? to 3 per cent? to 8 per cent? to .5 per cent?
3. Why is there less precision in the results of the dice example (Table 7) than in the coin example (Table 6)?
4. A retail gasoline station proprietor wished to obtain from his customers the following four types of information: The average mileage of cars per gallon of gasoline, the name of the manufacturer of the tires on the cars, the place of residence of the customers, the proportion of customers using premium gasoline. One of these types of information could be obtained only by the census method, one by a relatively small sample, one by a relatively large sample, and representative information on one of them probably could not be obtained either by sampling or census. Identify the four types of information according to the preceding description.
5. Basing your answer on the quoted paragraph at the bottom of page 80, and Table 9, discuss the question of whether the Federal Trade Commission's gross sample was representative of large and small chain stores.
6. What are the essential differences between uncontrolled sampling and controlled sampling? Between extensive sampling, selective sampling, inclusive sampling, and weighted sampling?
7. Given the following information concerning the 25,900 farms in five counties in 1940.

COUNTY	OWNERS		TENANTS	
	Sell Milk	Do Not Sell Milk	Sell Milk	Do Not Sell Milk
A	243	608	110	542
B	437	219	633	461
C	1,190	2,166	2,240	1,892
D	946	239	1,412	928
E	1,762	2,913	2,812	4,147
Total	4,578	6,145	7,207	7,970

- a) Set up the distribution of a sample of 400 cases to be collected from these counties by agents to obtain information concerning "the number and breed of cows used in dairy herds of farmers selling milk, and the average daily production of milk per cow."
 - b) Set up the distribution of a sample of 400 cases collected by agents to obtain information concerning "the difference in living standards, if any, between farmers who sell milk and those who do not."
 - c) Is the sample of 400 large enough in each of the preceding investigations, i.e., is the number of cases in the sub-groups sufficient to provide for proper operation of the principle of statistical regularity? Could the sample contain less than 400 cases? Discuss.
8. Suppose that an investigation by the sampling method were to be made of the extent of employment, unemployment, and part-time employment in a city of 500,000 population. The committee in charge would have to consider the following methods:
- A. Using schedules in the hands of agents
 1. Visit one house on each side of the street in each block of the city
 2. Select in advance representative blocks in the city and visit each house in those blocks
 3. Start from a common point with areas whose boundaries run out from the center like spokes in a wheel and instruct the agents to proceed within their areas until
 - a) They have secured 1,000 completed schedules
 - b) They have visited 1,000 houses
 4. Get names and addresses of unemployed from the local welfare bureau and names and addresses of employed from 20 leading employers. Visit the former to obtain data on unemployment and the latter to obtain data on employment and part-time employment.
 - B. Using mail questionnaires
 1. Address the occupant at number 13 of every series of 100 street numbers, for example, send a questionnaire to the occupant at 13 Englewood Ave., 113 Englewood, 213 Englewood, etc.

2. Address every tenth person in the telephone directory
 3. Address every twentieth person in the city directory
 4. Address the first 50 persons in each letter of the alphabet in the telephone directory.
- a) Discuss the likelihood of securing a representative sample by each method, (1) in the "A" group, (2) in the "B" group.
 - b) Rate the four methods of each group according to the kind and degree of control exercised in the sample.
 - c) How would you obtain this information by a completely uncontrolled sample?

REFERENCES

BROWN, LYNDON O., *Market Research and Analysis*. New York: The Ronald Press Co., 1937.

Chapter 10 deals with the planning of a sample.

HARPER, F. H., *Elements of Practical Statistics*. New York: The Macmillan Co., 1930.

The presentation of principles and methods of sampling in chapter I is very stimulating, particularly the discussion of weighted sampling.

Standards of Research. Des Moines, Iowa: Meredith Publishing Co., 1929.

Pages 27 and 28 give a specific statement of the policy followed in distributing questionnaires.

YULE, G. UDNY, and KENDALL, M. G., *Introduction to the Theory of Statistics*. London: Charles Griffin and Co., Ltd., 1937.

Chapter 18 contains an unexcelled statement of the theoretical background of sampling.

CHAPTER VI

COLLECTION OF DATA—DIRECT SOURCES

DESCRIPTION OF DIRECT SOURCES

IN CHAPTER IV direct sources were defined as the business concerns, government and private agencies, and individuals from whom statistical information not otherwise available could be secured by direct appeal. The type of source to which appeal will be made depends upon the kind of information desired. The internal records of business concerns are the original sources of such data as sales, profits, costs of doing business, employment, wages, and prices. Most of the information is private and can be obtained only on a confidential basis, but business concerns have come to realize the advantage to themselves and to the public of supplying the information for statistical purposes, provided the use made of it is not detrimental. Hence a large amount of valuable information may be obtained directly from business concerns.

The next sources from which information can be obtained are government and private agencies. Government agencies here refers not so much to those engaged in collecting and publishing statistical information as to those directly concerned with the control or regulation of business. Examples are the Board of Governors of the Federal Reserve System, the Federal Trade Commission, and the state public utility commissions. Agencies such as these are in a position to supply a great amount of information on special subjects in addition to that which they publish. Private agencies include trade associations, labor organizations, industrial institutes, charitable organizations, statistical services, research bureaus, and co-operative groups. In many cases these agencies are more valuable sources than business concerns. This is particularly true when information for an industry or an area is desired rather than for individual firms.

Finally, the statistician who is interested in data relative to consumption habits must expect to get his information from individuals or family groups. This is perhaps the most difficult source from which to obtain data because of the large number of persons who must be canvassed in order to get enough data for statistical purposes, and

because individuals so frequently do not possess the desired information or are unable to give it accurately even though it concerns themselves.

There is considerable difference in the actual collection process from direct sources according to whether or not the data exist in the files of the informant in such a form that they can be transferred to collection blanks. Collection from business firms and similar organizations quite commonly means merely transferring data from the records. On the other hand collection from individuals may require a lengthy process of interviewing to secure the information wanted.

COLLECTING DATA FROM DIRECT SOURCES

Once it has been determined that the collection of data must be made from direct sources, and the preliminary decisions concerning census or sample and the use of agents or mail questionnaires have been made, the investigator is ready to put his plan into operation. The work follows a natural sequence of steps regardless of the size of the investigation. The several steps are: (1) the provision for physical equipment; (2) a preliminary study of the field; (3) the choice of cases for the sample; (4) preparation of agents' schedules or mail questionnaires; (5) the selection and training of a staff; (6) supervising the work of collection.

There is literally no end to the amount of detail which might be introduced in discussing these steps. The intention is to present no more explanation than is necessary to give a broad view of the work. There is a wealth of reference material available from which more detailed information can be obtained.

The Provision of Physical Equipment

An office must be set up as a headquarters for the investigation. In some cases only pencil and paper and a place to work are needed. In general, however, equipment for filing, tabulation and calculation, typewriters, forms for recording the progress of the work, and similar materials must be provided.

A Preliminary Study of the Field

No matter how carefully the general plan of an investigation has been developed there will be certain peculiarities which need to be

discovered and provided for before starting the actual collection of data. A preliminary study may bring them to light and at the same time pave the way for the regular work. If there are technical terms used in an industry, these should be known in advance. A knowledge of the form in which records are kept and the units in which data are recorded will aid in phrasing questions. The advice of leading firms or agencies will be useful in showing the proper method of approach to others who are to be canvassed. This advice will be particularly valuable, if there are some concerns that are difficult to approach.

A common practice is to test a preliminary draft of questions by submitting them to a small sample of those from whom the information is to be obtained. The knowledge acquired in this way will aid in preparing the final draft of the questions, provide the background for improved agent technique, and create advance good-will for the investigation.)

The Choice of Cases for the Sample

! The method of selecting the cases to be included in a sample is one of the most vital steps in the entire collection process. For that reason all of chapter V is devoted to an explanation of the principles of sampling and the methods of choosing samples. The exact plan to be followed in selecting cases must be worked out in advance by the director and the importance of conformity to the plan must be impressed upon everyone connected with the investigation. If any subsequent change in the plan of sampling becomes necessary, such change should be made only with the knowledge of the director. For example, if a field agent in a consumer survey has been assigned a particular family, and finds it unwilling to give information, he should not try the house next door but should obtain a new assignment from the office. !

Preparation of Schedules and Questionnaires

The success of an investigation depends to a large extent upon the quality of the questions used. There will be considerable difference in the type of question included depending upon whether schedules in the hands of agents or mail questionnaires are employed. Agents can generally secure replies to questions which are more involved and more personal than those on mail questionnaires. In spite of this difference, the two types of lists of questions can best be discussed together with

separate explanations of the points that refer to one and not the other. There are four things to be considered in preparing a schedule or questionnaire: (1) content, (2) wording of the questions, (3) definitions, and (4) form.

Content.—In outlining the content of a schedule or questionnaire, the guiding principle is unity. The questions must be determined in terms of the objective of the investigation. Only those questions should be included which contribute directly or collaterally to the objective. Further, the questions must be so planned that the replies can be tabulated to yield answers to the questions proposed at the outset of the study. This requires that careful consideration be given to the ultimate goal of the investigation.

The object of a WPA project in an eastern city was to study the extent of the repair and modernization work which might be anticipated in the city under Title I of the Federal Housing Act of 1934. The schedule of questions in Figure 2 was prepared for the study. For multiple family dwellings a schedule was to be filed for each dwelling unit in the building.

FIGURE 2

SCHEDULE USED IN A REAL ESTATE SURVEY

1. How many occupants?
2. How many rooms?
3. Basement?
4. Stories?
5. Single or double garage?
6. Electric refrigerator?
7. Rent?
8. When was house built?
9. Owner or renter?
10. How long has occupant lived in house?
11. Automobile?
12. Use auto for work?
13. How long to go to work?
14. How many in family are working?
15. What kind of heat?
16. Fuel used?
17. Single or double house?
18. Is house in good condition?
19. Who pays water rent?

If questions 5 and 11 disclosed the fact that the family had an automobile and no garage, or a single garage and two automobiles, presumably

that family could be interested in garage construction. If question 15 showed that the house had no central heating system or had an antiquated system, perhaps the family would be interested in improved heating installation. If the answer to question 18 was a simple negative, further investigation of the house would be necessary in order to determine whether the deficiency was lack of paint, a rotting porch, a leaking roof, defective plumbing, or other needed repairs. Whatever the deficiency turned out to be, the family could presumably be interested in remedying it.

Some of the questions such as 6, 12, 13, 14, 16, and 19 are difficult to justify in this schedule; therefore in the revised form, Figure 3, they have been omitted. The proposed revision is designed to give more information concerning the repairs and modernization needed and to facilitate collection and tabulation. Agents using the revised schedule could save much time and effort because they would not be forced to ask any irrelevant questions and a better impression would be made on the informant.

FIGURE 3

PROPOSED REVISION OF REAL ESTATE SCHEDULE, FIGURE 2

HOUSE

- (1) Address
 (2) Stories: 1 2 3 4 B A (3) Single Double Other
 (4) Year built (5) Garage: 0 1 2 3 or more

DWELLING UNIT

- (6) Floor (7) Years lived in by present occupant
 (8) Owner Renter (9) Monthly rent
 (10) No. of rooms bath (11) No. of occupants
 (12) No. of automobiles owned
 (13) Heating equipment—central heat: Yes No
 (14) Hot air steam hot water
 (15) Any repairs needed: Yes No

REPAIRS NEEDED

- | | |
|-------------|---------------------|
| (16) House: | (17) Dwelling unit: |
| Paint | Electric wiring |
| Porch | Plumbing |
| Roof | Heating system |
| Sidewalk | Other |
| Driveway | |
| Garage | |
| Other | |

Wording of the Questions.—When schedules in the hands of agents are used as in the real estate survey of the preceding section, it is not necessary to word questions in so much detail as in a mail questionnaire. Since the agents are already familiar with the meaning of each question and the definition of terms, the abbreviated form used in Figure 3 is better than the sentence form of questions in Figure 2. It is easier for the agent to check the answer wherever possible than to write several words, and the uniform marking greatly facilitates tabulation. Where there may be a variety of answers a space is left for writing. For example, the answer to number 6 would be "whole house" for a single family dwelling.

The wording of a mail questionnaire is analogous to the agent's conversation with the informants. Since the questionnaire must be filled out by the respondent himself, the questions must be complete sentences and must make their own appeal. Certain practices in wording questions have been found so effective as to have almost the force of rules.

The wording must be clear to the respondent: Each question should contain but one idea. It must be stated as simply as possible so that there can be no doubt in the mind of the respondent what is wanted. Care should be exercised also to avoid the possibility of an ambiguous answer. For example, the following questions and answers are taken from an investigation made some years ago of the status of the Negro in industry.

In what jobs are both Negroes and whites employed? *Common labor*

In what jobs are only Negroes employed? *Common labor*

In what jobs are only whites employed? *Foremen and Mechanics*

These questions appear to be simple and straightforward. The investigator felt that there could be no doubt that they would be clear to employers to whom the questionnaire was sent. Yet impossible answers such as those recorded for questions 1 and 2 were received in a large number of the replies. Apparently the person filling out the questionnaire disregarded the word "only" in the second question. The maker of a questionnaire cannot expect that respondents will read questions as discriminately as was required in this case.

The work of the respondent must be kept to a minimum: There are several things to consider in complying with this rule. The use of a few easily answered questions in a questionnaire will increase the per cent of replies. If the answers can be given in a few minutes, the respondent is likely to fill them in immediately, whereas a list requiring more time may be laid aside and never picked up again. The number of replies is increased by the use of questions answered by "yes" or "no," or with easily obtained numerical answers or with a list of colors, qualities, places, etc., from which the respondent can check or underscore the applicable ones,

The respondent should not be asked to make computations. Hence the question, "What is your annual remuneration?" is not a good one to ask a laborer for two reasons. (1) Not only is the word "remuneration" foreign to his vocabulary, but (2) he may be unable to state his earnings except by the day or week. Requests for past information should be avoided if possible. The United States Department of Agriculture was not likely to obtain much usable information from the following request sent to farmers September 15, 1930:

	1929	1928	1927	1926
1. Acres sown to wheat in the summer or fall of each year.....	-----	-----	-----	-----
	1930	1929	1928	1927
2. Acres sown to wheat in the spring of each year	-----	-----	-----	-----
3. Acres of wheat harvested in the summer of each year.....	-----	-----	-----	-----
11. Actual cost of storing in local station elevator, per bushel, per month in each year	-----	-----	-----	-----

Make sure that no unnecessary repetition of information is requested. Two adverse results arise from failure to heed this warning: (1) The duplication adds to the length of the questionnaire and may be the cause of its being discarded. (2) The impression in the mind of the respondent created by the duplication is likely to be hostile to the point of causing him to discard the questionnaire.) Examples of repetitious and overlapping questions are found in the following list selected from a questionnaire sent to state hospitals by a social agency:

6. To what extent does overcrowding express itself in unsuitable sleeping quarters?
11. Do you need additional employees?
15. Do you have adequate hospital and medical facilities?

19. Do you have adequate facilities for giving your inmates instructive work and recreation?
 21. Are your facilities for academic work sufficient?
 22. (c) Is your staff of teachers large enough?
 28. Are inmates paroled whom you deem unfit to return to the community?
 29. What are the outstanding needs of your institution?
- | | |
|------------------------|-------------------------|
| Bed capacity | Medical equipment |
| Employees | Academic equipment |
| Teachers | Recreational facilities |
| Opportunities for work | Extended parole |

The last question merely asks for information already covered by the preceding questions. Questions 15 and 19 each combine two separate ideas. There are other faults in the wording of these questions which will be referred to later.

Form and content must not be offensive: A great amount of personal and business information can be obtained by the use of questionnaires, but great care must be exercised to avoid offense. One cannot ask the question, "What was the dollar value of your net sales last year?" But the approximate data may be secured by asking, "Please indicate in which of the broad groups below your net sales for last year would fall," followed by several sales classes arranged to give enough detail for use in the subsequent steps of the investigation. A question may not be personally offensive, but may involve official complications. In the example quoted in the preceding section, a hospital superintendent might well hesitate to answer question 28, fearing to give offense to the parole board and to politicians. A question stating or even implying moral turpitude should be avoided. Likewise questions dealing with religious principles or habits should be used with caution.

Bias must be avoided: Bias may enter in two ways. First, the question may be phrased so as to lead to a certain answer. An example of a biased wording is, "Did the fish cakes taste better to you than canned salmon, salted cod, or shredded cod?" It would be much better to list the four types of prepared fish and request that the user number them in the order of preference.

Second, estimates that are based on opinions rather than on actual figures may be biased. Suppose you were inquiring of a manufacturer of drugs whether his product was distributed at retail mainly through chain stores or independent stores. His direct contacts with the buyers of chain retailers might lead him to suppose that they were his chief

customers, whereas a study of the sales records might well show the reverse.

Answers should be obtained in the most usable form: This is essentially a matter of visualizing the subsequent use of the data. In particular all units should be carefully selected and defined from the point of view of the subsequent analysis. When new information is to be used in conjunction with some already in hand be sure that the new information will be comparable with the old.

It is also essential that the information be received in a form which facilitates tabulation and analysis. In most questionnaires having more than two or three questions cross-information becomes available by using the results of two or more questions together. It is important to plan the questions so as to develop the maximum amount of cross-information. Conversely, failure to consider this feature of the questionnaire may lead to a serious hiatus in the information which will be discovered too late to be remedied. Figure 4 shows the advantages of foreseeing the subsequent parts of the work when making the questionnaire.

The three blank spaces marked "Leave blank for Department use" permit the tabulation on this form of the number of spindles in the mill, the number being currently operated, and the number belted and ready to operate (active spindles as defined in the industry). Information is also available from which to compile an age distribution of total spindles and active spindles in the mill. Likewise a distribution by kind of spindle can be made for total spindles and active spindles and these can be further classified by age, if desired. Thus we see the amount of information that can be taken from a carefully prepared questionnaire such as this one. All of the tabulation forms were drawn up at the same time the questionnaire was prepared and the two were made to conform at every point.

Definitions.—In preparing a schedule or questionnaire, any word, phrase, or technical term which may lead to variation of interpretation should be defined. The units in which the data are to be collected must also be defined. These definitions are equally important in the case of schedules and questionnaires, and the necessity for preliminary decisions regarding the terms to be used is the same in either case. However, it is evident from the rules regarding questionnaire construction that if a great amount of detailed definition proves necessary the use of the mail questionnaire is to be avoided. Whatever defini-

FIGURE 4

WORSTED SPINNING SPINDLE INVENTORY, MARCH 1, 1931

A. *Worsted Spinning Frames and Mules in Mill No.* , *March 1, 1931*
 (Use a separate schedule for each worsted spinning mill)

Please list the following information about all worsted spinning frames and mules in the possession of this mill on March 1, 1931. (Include spindles producing for firm's own consumption as well as spindles producing for the sales yarn market.) You may either list the information separately for each machine, or according to the lots in which you purchased them. For instance, a lot consisting of 10 cap frames—200 spindles each, bought second-hand in 1919, of which 5 were not operated at all in 1930, and of the lot, 7 are now in place with belts on—would be listed as indicated on the first line of the form below. Where a lot contains more than one kind or size of machine, please use a separate line for each kind or size.

Note: The Department pledges that all figures returned on this form will be kept strictly confidential.

Number of machines in the lot	Check kind of spindle below				Number of spindles per machine	Leave blank for Dept. use	If bought prior to Jan. 1, 1914 check below	If bought since Jan. 1, 1914:			*How many machines in this lot were not operated any in 1930?	Leave blank for Dept. use	How many machines in this lot are now in place with belts on?	Leave blank for Dept. use
	Cap	Eng. Ring	Flyer	Fr. Ring	Wstd. Mule			Give year of purchase	Check	Second Hand				
10	✓							1919		✓	5		7	

* If this worsted spinning mill was completely idle during the entire year of 1930, when was the last year of any operation

tions are essential in a questionnaire should be printed as close as possible to the questions to which they apply. The definitions usually do not appear on a schedule but should be printed along with the general instructions to the agents.¹

Terms: For either method of investigation, the definitions must be inclusive of all the limitations that have been placed upon the collection process. They must be so precisely worded that (1) no ambiguity of terms exists; (2) no limitations of terms are left indefinite; and (3) no technical uses of terms are unexplained. Some examples will show the necessity for careful wording and definition;

The treasurer of a department store submitted a list of questions to the department heads of the store. One of them read, "Have you been successful recently with promotions?" There may be no doubt as to what is wanted here, but the simpler thing would be to specify sales promotions rather than promotions of the staff.

Referring to the questionnaire sent to state hospital superintendents (p. 98), question 6 reads, "To what extent does overcrowding express itself in unsuitable sleeping quarters?" The phrase "to what extent" is indefinite. Such words as "unsuitable," "adequate," and "sufficient" as used in this questionnaire are meaningless unless related to definite standards.

A questionnaire sent to college and university teachers contained this question, "What per cent of your regular salary goes for rent?" This question appears to be simple but the word "regular" is used in a technical sense. Presumably it means the compensation for teaching the usual number of hours per week for a nine months' period. This definition would exclude evening and extension school salary in some cases and include it in others. Summer school salary would not be considered "regular" even for a person who taught every summer. Many other variations exist in different colleges. The word "regular" requires exact definition if the results obtained from the questionnaire are to be comparable.

Units: Every kind of recording of numerical information requires that a unit be established in which to perform the process of enumeration.¹ The unit may be a person, an animal, an inanimate object such as a tree or a house, a measured quantity such as a ton or a bushel, a money measure such as the dollar or the franc, an abstract concept such as an order, an accident, or a vacation. In some cases the type of enumeration to be made immediately determines the unit to be used,

as in counting population or recording sales. In other cases a choice of units is available as in recording production of cement in which the count could be made in tons, barrels, or dollars of value. In every case in which the unit is not obvious from the nature of the enumeration, selection of a unit must precede the counting process.

Once the unit has been selected consideration must be given to the question of whether any uncertainty may arise in its use. In many cases definition will be required to avoid ambiguity; thus in collecting information concerning the size of houses careful definition of what to count as rooms is necessary. Similarly, in recording industrial accidents, a careful statement must be made of what sorts of injuries are to be included as accidents.¹

Units can be divided into two kinds: (1) those with definition established by law or custom, and (2) those for which the definition must be established separately wherever they are used.¹ Examples of the first kind are the bushel, the gallon, the yard, the hour, etc. Each of these measures carries a standard definition which serves as an adequate description any time it is employed as a unit.¹ The unit for measuring wheat is the bushel, and no further definition is required. On the other hand when the unit used is a ship, a room, a voter, or a horse it is necessary to explain what shall be counted and what shall be omitted during the enumeration.¹ Thus a room in a dwelling is not a usable unit until many borderline cases such as closets, breakfast nooks, pantries, and sun-rooms have been either included or excluded by definition. If a subsequent investigation is made using a different definition of a room, the results of the two investigations cannot be compared although both use the unit "room" as the basis for counting.

These two types of units are usually known as measurement units (fixed definition) and counting units (variable definition). To call the latter type counting units seems somewhat ambiguous because the process of enumeration involves counting regardless of the type of unit used. For that reason we prefer the distinction based on the amount of definition required.¹

Variable definition units such as a "room" exist independently of the counting process, a fact which explains the need for definition each time they are used. As separately existing entities they possess individual differences which inevitably run to limits where it is necessary finally to establish a boundary. The same situation does not arise when a unit such as the pound or mile is used.

Definition is required to different degrees in the realm of variable definition units. A "person" needs little or no definition because the unit is so universally recognized. Likewise the unit "citizen," a part of the universe of persons, has a fixed legal definition in each country. On the other hand such a unit as a "salesman" or a "criminal," each a part of the universe of persons, requires very careful definition. Only a few units are so well recognized as the person or citizen, hence the general conclusion that one must always be prepared to state variable definitions completely. An example will illustrate what is likely to occur in particular cases.

The instructions accompanying a schedule included this statement: "Information to be secured for wage-earners only." Trouble arose continually in determining exactly who were wage-earners. The general concept "wage-earners" was perfectly clear, but the definition of the statistical unit a "wage-earner" was extremely difficult. On the face of it a wage-earner should be one who receives compensation from others for services rendered. But questions such as the following were brought in by the agents daily: "How about a physician who received a salary instead of fees?" "How about a daughter living at home and working for her father at a nominal salary?" "How about an insurance agent working on commission and receiving a fixed percentage of the annual profit?" As soon as these questions were answered, others arose. The answers given to such questions as these depend upon the purpose of the investigation. But no matter how carefully any such unit as "wage-earner" is defined borderline cases will arise which will have to be settled arbitrarily by the director.

Form.—General: The primary consideration that determines the form of a schedule is convenience, whereas the first requisite of a mail questionnaire is good appearance. In either case the size, shape, material, and type of printing are important. Whenever possible cardboard rather than ordinary paper is to be preferred. Cards are easier to handle in an interview and more durable for editing and tabulating. If cards are not feasible then a good quality paper should be used. One sheet of questions is always preferable. However, if the choice is between overcrowding of questions and the use of an additional sheet, the latter is the lesser of two evils. Overcrowded questions are likely to result in misplaced answers, incomplete answers, and increased difficulty of tabulation.

The first impression made by a mail questionnaire will determine to

a large extent whether it will be answered or discarded. A closely printed sheet or card immediately gives the impression of being lengthy and time-consuming. This can be partially overcome by well-spaced questions, the use of rulings, and variations in type sizes.

Sequence of questions: The questions should be arranged so that they will form a natural sequence for the respondent. Although the sequence must be varied to meet the requirements of different investigations and will not be identical for schedules and mail questionnaires some general principles can be stated.

1. The initial question or questions should be simple.
2. Any preliminary questions which are necessary to pave the way for the key questions should come next.
3. The key questions, that is, those which relate to the major purpose of the investigation, should be placed at the end or near the end.
4. If one or more questions calling for an opinion rather than a statement of fact are included in the schedule, they are usually placed at the end.

Figure 5 taken from the questionnaire used early in 1935 in a market survey of college men employs a good sequence with one exception.

FIGURE 5

RADIO SECTION OF QUESTIONNAIRE USED IN SURVEYING THE COLLEGE MARKET

1. Do you have a radio in your *college room*? Yes No
2. What make?
3. When bought? 4. How many tubes?
5. Where bought? College town Outside college town
6. Do you intend to purchase a radio before 1936? Yes No
7. If so, what make? 8. About what price?

Questions 1, 2, and 4 provide the facts concerning the student's present radio. The questions are simple and the information is immediately available. Questions 3 and 4 should exchange places. Questions 3 and 5 give preliminary information leading up to question 6. Question 3 was presumably intended to give information concerning the age of the radio. The answer will, of course, be misleading if the student bought the radio second hand. Question 6 is the key question of the schedule. It gives the information concerning the potential market for radios among college men during the current year.

Questions 7 and 8 correctly follow question 6 since they ask for information supplementary to it.

Auxiliary material: In addition to the list of questions certain explanatory material must be prepared. Most important is a letter of transmittal to accompany a questionnaire or instructions to agents who collect schedules. The purpose of the letter of transmittal is to engage the attention of the addressee and encourage him to respond. The instructions to agents provide a background of information which will permit them to accomplish the same result by personal interview.

Letter of Transmittal with a Questionnaire—Motives: In a mail questionnaire the questions themselves may be preceded on the same sheet by a brief explanation of the purpose of the investigation and the reasons for requesting information from the particular persons to whom the questionnaire is sent. The usual method, however, is to inclose a letter of transmittal explaining the questionnaire and pointing out some incentive for answering. When sent to business men it is also customary to include a duplicate copy of the questionnaire for the respondent's own files.

Some of the motives to which an appeal for voluntary replies can be made are: (1) co-operation; (2) interest; (3) profit; (4) obligation; and (5) position. Figures 6–10 illustrate from actual examples how appeals can be made to different motives.

FIGURE 6
CO-OPERATION

DEAR SIR:

Will you be kind enough to take a few moments of your time to jot down the answers to the questions on the back of this letter?

We would greatly appreciate this favor. You will not be asked to buy anything—your name will not be used at all.

This letter is one of a small number we are sending out at the request of a large manufacturer to get the viewpoint at first hand from some of his customers.

It is not necessary to write a letter. Just check or fill in your answers in the space provided and mail the sheet back to us in the enclosed stamped envelope.

Many thanks for your help.

Very truly yours,

[An Advertising Agency]

FIGURE 7

INTEREST

DEAR FELLOW-MEMBER:

The subject of calendar reform has been studied during 1933 by a committee of the American Statistical Association, and its majority and minority reports appear in the supplement to the Association's Journal for March 1934. Before taking action, the Board of Directors wished to pursue the subject further, and has appointed the present Committee, with instructions to ascertain the considered opinion of the Association's membership on the question of calendar reform.

The problem presented by the defects of the present calendar is obviously of importance to the statistical profession, a number of whose members deal with the analysis of time series.

It is the purpose of this Committee to obtain as far as possible the considered opinion of the whole membership.

To aid in this undertaking, will you kindly fill out the questionnaire and return it to the Committee at the earliest possible date.

Yours very truly,

[A Committee Chairman of the American
Statistical Association]

FIGURE 8

PROFIT

DEAR SIR:

In order to assist farmers in adjusting the production of milk and dairy products to prospective demand, the U. S. Department of Agriculture is now undertaking to collect more complete information regarding the number of cows being milked, the quantities of milk and cream being produced and sold and such information regarding the number of heifers being raised, the number of cows coming fresh, the quantity of grain being fed and dairymen's plans for the future as may be needed to find what changes in production may be expected.

Those who cooperate with the Department by returning each month a report for the herd which they own or operate will receive copies of the reports.

On the other side of this page you will find some questions regarding the quantity of milk now being produced on your farm, the last price received, and the quantity of grain being fed. In return for your assistance I am enclosing a summary of the outlook for dairying so far as this Department has been able to determine the outlook from such information as is now available.

Yours very truly,

[A Division Chief of the United States
Department of Agriculture]

FIGURE 9

OBLIGATION

DEAR SIR:

Those industries which today are making progress toward stabilization know their capacity as well as their demand. To bring some light to the capacity situation in worsted spinning, the Research Department reported briefly to the trade in December on post-war trends in worsted sales yarn spindles. Lack of figures at the time of that release prevented a detailed analysis of current worsted spindlage. We are now ready for that step—an industry-wide inventory of all worsted spinning spindles in the textile mills of this country as of March 1st, 1931. Leaders of your industry have approved this survey. Typical of their attitude is that expressed in a recent letter received from Mr. _____, a copy of which we are glad to submit at this time.

You are well aware, of course, that a survey of this nature is only successful to the degree that all firms in the industry respond. In a word the value of this survey to you is very directly tied up with the number of firms in the industry who supply the information outlined on the enclosed schedule. An early reply on your part will help insure an early report of the results by this Department. Please do not hesitate to bring to our attention any problems that may arise in filling out this schedule. We will gladly assist you in any way possible.

Sincerely yours,

[A Director of a Research Bureau]

FIGURE 10

POSITION

FELLOW-ECONOMIST:

In order to help toward straight thinking on the prohibition question, will you please fill in the accompanying questionnaire to the best of your ability?

Please note that many of the answers must represent opinions, but that your unbiased judgment *as an economist* is desired.

I assure you that your name will not be mentioned in any way unless you give permission.

Please do it now, using the enclosed stamped envelope for mailing your reply.

Sincerely yours,

[A University Professor]

Distinct from questionnaires which depend upon making an appeal to some voluntary motive are government requests for information to which answers are compulsory. The element of compulsion is predominant in such a letter, as Figure 11 will show.

FIGURE 11
COMPULSION

SST-1564-CLH

TREASURY DEPARTMENT
Office of the Collector of Internal Revenue
Pittsburgh, Pa.

July 29th

..... 1937

B..... Company

M....., Pa.

The records of this office indicate that you filed an application on Form SS-4 for an "Employer's Identification Number" under the provisions of Title VIII of the Social Security Act and that you were assigned the Identification Number above indicated. This Act requires every Employer to file a return for each month and pay the tax shown to be due, effective from January 1, 1937.

The records of this office indicate that you have not complied with the law in this respect by reason of your failure to file a return for the months of January, February, March, April, May, and June, 1937, inclusive.

You are, therefore, requested to file a separate return on the blank forms inclosed for each month above mentioned and to forward the same to this office with your remittance for the tax due. An affidavit in explanation of your failure to file the returns within the time prescribed by law must accompany the returns for the consideration of the Bureau in connection with the assertion of the delinquency penalties. (Blank affidavit inclosed.)

In preparing these returns, your complete Name, Address and your Identification Number must be shown thereon as indicated at the top of this letter. If, for any reason, you are not subject to the provisions of this Act, please advise this office fully, or, if a return was filed by you in another District, advise the date and the place where filed, also the serial number stamped upon your cancelled check.

Reply should be made within ten (10) days from the date of this notice.

Very truly yours,
[An official in the Internal Revenue Office]

Sometimes several motives for securing replies will be combined. The question of what type of appeal to use will depend upon the particular circumstances involved. Considerable care should be exercised in writing these letters because they must accomplish in an investigation by mail what an agent does by personally interviewing prospective informants when schedules are used.

Instructions to Agents: There are two parts of the instructions to agents: (1) the definitions of terms and units used in the schedule and (2) the general instructions. The definitions were discussed on page 100. An illustration of what should be included under general instructions is provided in Figure 12, which was prepared for a study of home ownership in Buffalo, New York.

FIGURE 12
INSTRUCTIONS TO COLLECTING AGENTS
PRELIMINARY

You have in your possession a letter identifying you as an agent of the President's Conference on Home Building and Home Ownership. Use this identification discretely, remembering always that you must not use compulsion in seeking replies; on the other hand, you do have back of you the authority of a government investigation which should instil confidence and insure to the informant that all information given will be treated as absolutely confidential. In the use and publication of results no individual information will ever be divulged.

No identification appears on the schedule except the case number. On the separate sheet which is provided for the purpose keep an accurate record of the exact address from which each case is taken. Agents must be doubly careful to keep this special record accurate or the editing of the questionnaires may be seriously handicapped.

This study is confined to families having total income (earned or other) not exceeding \$3000. You will be unable to determine exact income before reaching page 3 of Form 1. A preliminary question, however, should determine the availability of the family. If exact tabulation of page 3 should show total income slightly in excess of \$3000, the schedule will be used.

Information is to be collected only from families composed of a minimum of husband, wife, and one dependent child.

Where you find more than one family occupying quarters which are quite clearly intended for a single family, do not fill out Form 1 but secure the information on Form 2.

Information is to be collected only from families who were purchasing homes during 1930, but who began paying for their homes prior to January 1, 1930.

Information is to be collected only from families in which both parents are native born whites.

The most inexcusable errors in compiling data are those arising from carelessness on the part of collecting agents. Therefore,

- (a) Write legibly; be neat.
- (b) Be sure that you understand all questions and instructions.
- (c) Before closing your interview, check to avoid omitting any part of of your schedule.

Selection and Training of Staff

Types of Workers.—The number and types of workers needed depend entirely upon the character of the investigation. If it is a questionnaire study, the problem of organizing a clerical staff for the preliminary work is in no way peculiar to a statistical inquiry. However, when agents are to be used the selection of personnel presents a more specialized problem. In both types of investigation, editors and a staff of statistical clerks will be needed as soon as collection is under way. The selection and training of all these workers becomes a part of the process of conducting an investigation, particularly if the only available staff consists of students or other inexperienced workers.

Qualifications and Training.—It is essential not only that each of these workers receives instruction as to his own specific duties, but also that each acquires a thorough understanding of the entire process.¹ For example, if the agents receive some training in the method of tabulation, they will realize why they are asked to make entries in a uniform manner. If the editors have even a slight experience in collecting the schedules, they understand the difficulty of getting exact information and will be more likely to offer their criticisms to the agents without arousing antagonism. Testing of the various staff members in different parts of the work has the additional advantage of discovering which ones are best adapted for editing and which ones make the best agents.

Some individuals may not be successful in making personal contacts and getting information from other people but may have an eye for detail and a capacity for detecting errors. The latter qualities are desirable in an editor, but even more important is the ability to appraise a schedule as a whole for consistency and validity. This is especially true when long and complicated schedules as in a cost-of-living inquiry are being collected. A schedule may balance perfectly having no specific errors of any kind and yet contain gross inconsistencies or omissions that can be detected by an editor with common sense.² For example, in a cost-of-living investigation good editing would immediately question a schedule in which some income was reported from insurance benefits resulting from the death of a member of the

immediate family, but where no item for funeral expenses appeared under "expenditure."

Agents become the direct representatives of the organization conducting the inquiry in making contacts with those from whom the data are to be obtained. Upon their shoulders rests to a considerable degree the responsibility for the success of the undertaking. A good agent must have "tact"—defined as "intuitive perception, a ready appreciation of the proper thing to say or do, especially a fine sense of how to avoid giving offense." In other words he must be a good salesman in order to "sell" to the informant the idea of answering the questions.

Training in certain techniques must be added to this natural capacity before the agent becomes a qualified field representative. First, he must thoroughly understand the general purpose of the investigation and believe in it himself. He should be able to explain it and convince the informant of its validity without the necessity of referring to his letter of credentials. Agents must always be furnished with such credentials for their own protection, but an official letter practically never has a persuasive effect on an irate informant even though the investigation is being conducted under government authorization. Next, the agent must be so familiar with the schedule and all of the instructions, definitions, and limitations that he can conduct the interview and complete the schedule without any hesitation or reference to notes. He is never permitted to alter the meaning of a question or definition and if in doubt on any point he should add full notes describing the situation. If an unusual situation not contemplated by those who planned the investigation should arise, a well-trained agent should be prepared for such contingencies. His function is to secure complete information on the unusual case so that final disposal can be made by the person in charge of the investigation. In such cases it may be advisable not to complete the interview, but to leave the way open for a return visit after having consulted the director for advice.

All of the information which the agent secures, whether written on the schedules or given orally in an interview, is completely confidential. Inexperienced agents sometimes forget that they are not at liberty to discuss collected information with anyone, even with fellow-agents and much less with friends. Failure to observe this rule can be extremely embarrassing, if information that was given in confidence comes back to its origin through third parties, and may have even more serious consequences.

Example of Agent Technique.—The following example shows the effect of proper and improper agent technique:

Several years ago an investigation of housing conditions was being made in a large city. Although the investigation was sponsored by a committee which had been appointed by the President of the United States, it was not official nor was anyone compelled to give information to the agents. In spite of the fact that the status of the investigation had been explained fully, one of the agents insisted upon answers to his questions to a point which caused an irate housewife to call the police station. The agent was picked up by a policeman on the housewife's complaint, taken to the station house, and subsequently to the office of the police commissioner. At this point the director was called to the commissioner's office. The director's explanation convinced the commissioner that no crime had been committed, but did not convince him that agents should be permitted to annoy housewives any further. Fortuitous circumstances entered the case at this point. A detective who had been assigned to the case was called in by the commissioner. Quite unexpectedly the detective reported that another agent had appeared at the door of his home the previous evening, that the agent had been courteous, his questions inoffensive, and that the detective had been entirely willing to give the information requested. The commissioner then consented to have the investigation continue provided the offending agent were dismissed.

Perhaps it is superfluous even to point out the value of the second agent's work. By the proper approach this man had secured the information that he wanted, had created good-will for himself, and quite unwittingly had saved the entire investigation. In addition to showing how agents may succeed or fail, the example points to the desirability of notifying police authorities before sending agents out to do house-to-house investigation.

Supervising the Work

The foregoing example indicates the necessity for constant supervision on the part of the director while the investigation is in progress. He must be available at all times so that any unusual situations can be met as they arise. On the other hand every detail of the routine plan should be a matter of written record, and each part of it should be thoroughly understood by at least one other staff member, so that the orderly progress of each step is automatic regardless of the presence

or absence of any one person. These routine steps include: the checking of the adequacy of the sample as the collection proceeds; a regular system for making assignments, accounting for returns and routing of schedules to agents, to editors, back to agents if necessary, and finally to tabulator; arrangement for check interviews by visit or telephone as a test of each agent's ability and integrity; adherence to the quota and time schedule originally planned for the investigation; issuance of additional or revised instructions to the entire staff whenever necessary; and provision for staff meetings at regular intervals for the discussion of difficult points that may arise. /

SUMMARY

In the conduct of any specific investigation numerous details arise that cannot be discussed in a general textbook. No attempt has been made in this chapter to furnish a complete guide for a person undertaking a statistical investigation in any particular field. Many books¹ devoted solely to the description of research techniques can be consulted to supplement the statement of principles and methods presented here.

PROBLEMS

1. Assuming that information on the following subjects is to be obtained by direct collection, which of the three types of sources listed in the text should be used in each case?
 - a) The brands of bread used by families in a city.
 - b) The distribution of employed persons in a city according to a classified list of occupations.
 - c) The extent of use of different types of anti-freeze solution in automobile radiators.
 - d) The extent of unemployment of union labor and non-union labor.
 - e) The tendency toward the construction of lower-cost houses in urban centers.
 - f) The distribution of vacant dwellings in a city according to rent level.
 - g) The effect on sales of milk in a community of an increase of 7 per cent in the retail price.
2. What is the purpose of preliminary testing before starting a direct investigation?

¹ A few of these are listed at the end of the chapter.

3. Explain the difference in wording of questions in a schedule and a questionnaire.
4. Explain which of the following alternative wordings is preferable for a questionnaire and why:
 - a) (1) What color do you prefer for your next automobile?
 - (2) Check the color you prefer for your next automobile:

green	maroon
blue	brown
grey	gun metal
black	other (specify)
 - b) (1) Do you consider the advertising statements of local stores more reliable than statements found in magazine advertising? More
Less
Do you consider the statements made in advertising over the radio more reliable than those found in newspapers? Yes No
Do you feel that a statement in an advertisement is more reliable than a statement by a clerk in the store? Yes No
 - (2) Mark in the order you consider them dependable the following media of information concerning consumers' goods (mark the most dependable 1, etc.)

magazine advertising
radio advertising
newspaper advertising
statements of clerks in stores
 - c) (1) Do any of the following apply to your concern? (Check which.)

too many salesmen
sales management inefficient
sales territory poorly allocated
sales commissions too large

 - (2) Which of the following would be most effective in reducing selling expenses in your concern? (Check one.)

reduction of selling force
reorganization of sales management
reallocation of sales territory
reduction of sales commissions
5. Define the following terms for use in a schedule or questionnaire. Be sure to provide for possible borderline cases. a) a farm; b) a factory; c) an employed person; d) a department store; e) a radio news broadcast.
6. Explain the difference between fixed definition units and variable definition units. Give three examples of each not taken from the text. Explain the need for definition in each of your examples of a variable definition unit

- 7. Write a letter to accompany the radio questionnaire of Figure 5, page 105.
- 8. The following questionnaire was sent to the subscribers of a magazine by the management of the magazine. Write a letter to accompany the questionnaire.

Name..... Age.....
Address..... City.....
Occupation
Name of Company..... Position.....
Are you the head of a family? Number in family
Do you own an automobile? Make Year
Do you own your home? Number of Rooms
What are your hobbies?
Where do you spend your vacations?
Do you own a radio? Make
Have you a telephone?
Suggestions
.....
.....

- 9. Summarize the qualifications of a collecting agent.

REFERENCES

BOWLEY, ARTHUR L., *Elements of Statistics*. London: P. S. King and Son, Ltd., 1920 (fourth edition).
Chapter III presents rules for the preparation of schedules and several examples of direct collection.

BROWN, LYNDON O., *Market Research and Analysis*. New York: The Ronald Press Co., 1937.
Chapter 9 contains an excellent statement of the rules to be observed in preparing a questionnaire.

DAY, EDMUND E., *Statistical Analysis*. New York: The Macmillan Co., 1925.
Statistical units are discussed on pages 17–23.

EIGELBERNER, J., *The Investigation of Business Problems*. New York and Chicago: A. W. Shaw Co., 1926.
Chapters IX, X, and XII deal with the mechanics of direct collection.

SAUNDERS, ALTA G., and ANDERSON, CHESTER R., *Business Reports*. New York: McGraw-Hill Book Co., Inc., 1929.
Chapters VIII and IX deal with the collection of data; emphasis on the preparation of questionnaires.

Standards of Research. Des Moines, Iowa: Meredith Publishing Co., 1929.

Pages 20–26 contain an outline statement of the "Planning of Questionnaires." The use of agents is outlined on pages 40–43.

The Technique of Marketing Research. Prepared for the American Marketing Society by the Committee on Marketing Research Technique. New York: McGraw-Hill Book Co., Inc., 1937.

This entire book should be required reading for every person who expects to engage in marketing research.

CHAPTER VII

EDITING AND PRELIMINARY TABULATION

TWO IMPORTANT steps, editing schedules and preliminary tabulation, follow the collection of data and precede the preparation of tables for presenting the collected information. Inadequate attention has sometimes been given to these processes because they do not require the use of involved techniques. Nevertheless an understanding of the methods of editing schedules and of transferring information from schedules to the initial tabular form is an essential part of the conduct of an investigation. The two processes are distinct, hence this chapter has been divided into two parts.

EDITING SCHEDULES

As agents' schedules and mail questionnaires are returned they must be studied very carefully in order to detect any irregularities in the responses. Experience demonstrates that this step is necessary whether the collection has been made by agents or by mail, although more questions will be answered incorrectly in mail questionnaires than in schedules collected by agents. Before any analysis is undertaken these errors must be detected by an editor and corrected if possible.

The editor performs two functions: (1) detecting irregularities in the replies and (2) preparing the schedules¹ for tabulation.

Editing for Irregularities

There is no fixed order in which the editing should proceed. That is within the discretion of the editor. The following order will serve in many cases: (1) look for omissions, (2) verify check questions, (3) check for inconsistencies, (4) search for errors, and (5) check for uniformity between schedules.

Look for Omissions.—Each schedule should be complete. If the answers to any questions are missing an attempt should be made to get the information either by mail or by a second interview with the

¹ As used in this chapter the word "schedules" refers to collected information whether obtained by agents or through the use of mail questionnaires.

informant. Failure to obtain the information by these means may cause the editor to mark that part of the schedule "no report" or if the missing information is primary, to discard the schedule. In the chain-store inquiry cited in the chapter on sampling (pp. 79-81), 195 schedules were discarded entirely because primary information was missing and 50 other schedules were incomplete in some respect.

Verify Check Questions.—If the collection form includes answers to questions which should verify or check each other and these fail to check, the editor must search for collateral information that will indicate which of the responses is in error. For example, the age of a house may be stated as 22 years (in 1936), the date of construction as 1920, and the initial mortgagee as a bank which was liquidated in 1916. The date of construction has apparently been given incorrectly, but the editor must not guess about this. If no collateral verification is possible, either the schedule must be returned to its author or the answers to these questions must be discarded.

Check for Inconsistencies.—There will often be questions the answers to which can occur only in certain combinations or in certain sequences. The editor must test these combinations and sequences for consistency. Replies to the following two questions sent in by a gasoline and oil service station are inconsistent:

What disposal do you make of bulk motor oil distributed to you on quota and unsold? (check which)

Throw away ☐

Sell at lower price ☐

Return to agency for credit ☒

Mix with new oil received ☐

Allow to accumulate for waste use ☐

At what price per quart do you sell motor oil?

Heavy body 35 cents

Medium body 30 cents

Light body 25 cents

Old stock left over 20 cents

(known in the trade as last year's oil)

If last year's oil which is unsold is returned to the selling agency for credit, then the 20-cent selling price quoted has no meaning. Either

some old oil is sold or the price quotation on it should be removed. The editor must find out which answer is to be amended.

Search for Errors.—Any calculations which are on the schedule should be carefully checked. A tabulation of a total and its parts should also be verified. Errors which occur in numerical relations can usually be corrected by the editor. There are some cases, however, in which errors of this kind will require a resubmission of the schedule to the maker.

Beyond these obvious things there may be others which can be detected only by a careful study of the answers to all of the questions. For example, a research bureau was receiving monthly reports of sales from a number of department stores. The sales of one store seemed to move opposite to the others in June. After this had happened for the third year, the director of the bureau grew suspicious. The difference might be the result of a special sale in June, or be due to the handling of special seasonal merchandise, but preliminary study of the case failed to disclose any reasonable cause for the difference. Finally a direct appeal to the co-operating store disclosed the fact that the controller, who regularly made out the monthly report, took his vacation in July and his substitute who made out the report of June sales misunderstood the schedule and reversed the May and June figures.

Check for Uniformity between Schedules.—The editor should check for uniform interpretation of all of the questions. He is quite likely to find that one or several questions have been misconstrued on some of the schedules. These things may not be evident in studying the schedules individually but may appear when one question is studied on all of the schedules. In an investigation of moving picture attendance by students, this question was asked: "How much did you spend for moving picture admission last week?" One of the student agents was noted for his erratic behavior and on this question he ran true to form. Many of his schedules showed an expenditure well above that of other schedules. Inquiry elicited from him the fact that he had probably asked for expenditures during the past month. All of his schedules were dropped from the investigation.

Re-editing

If the investigation is a large one or the schedule form is complex, two editors should go over the schedules independently. The second editor will find things which the first one overlooked. In fact the

schedules will always be somewhat less than perfect. As much time and money as possible should be used to improve them. Sometimes it will be better to have different editors check different parts of the schedule. This plan has the advantage of concentration but is weak in that no one person gets a comprehensive view of the schedules as a whole.

The work of each editor should be distinguishable either by the use of different colored inks or other distinctive marking so that any corrections or alterations by an editor can be referred back to their author if necessary. It should be a fixed rule that editors do not erase; they cross out and substitute in all cases.

Preparing the Schedules for Tabulation

After the various irregularities have been adjusted, some steps still remain to be taken before the schedules are ready for tabulation. In the course of these adjustments changes will have been made on some schedules, unusual markings will appear on others. The editor should indicate specifically how such items are to be tabulated if there is any chance of subsequent misunderstanding. To facilitate the transfer of information from editor to tabulator, all final corrections should be made in ink of a certain color.

Finally, the editor should indicate the proper classifications of items if they are to be tabulated in a form different from the way they appear on the schedule. For instance, if the question "What is your occupation?" appears on the schedule and no check list accompanies it, the answers will appear in a variety of forms. The editor must mark these replies according to the occupational classification to be used in tabulation. Again, the schedules may show the state from which they come, whereas the tabulation is to be made by geographic areas. These areas should be marked by the editor. This sort of editing adds to the speed and accuracy of the tabulation and makes it unnecessary to employ a highly skilled staff in the tabulation process.

Sometimes the process of preparing the schedules for tabulation involves an intermediate step known as coding. The best example of this occurs when mechanical tabulation is employed. Coding of information to be transferred to "punch" cards becomes one of the most important steps in the mechanical process which is described in the second part of this chapter.

PRELIMINARY TABULATION

The methods to be used in transferring information from the collection forms to preliminary tables will depend upon the size of the investigation, the character of the data, and the ultimate form in which the results are wanted. The following four methods are available: (1) sorting-counting, (2) the use of a tally sheet, (3) the use of a work sheet, and (4) mechanical tabulation.

Sorting-Counting

The sorting-counting process can be used to advantage when the data to be tabulated are relatively simple so that each case can be put on one small card. The cards can be sorted and sub-sorted into piles according to any desired plan of classifying the data. The number of cards in each pile can then be counted and recorded on a tabular form prepared for the purpose.

The card used in an investigation of vacant dwellings in Buffalo, New York, has been reproduced as Figure 13. The purpose of the investigation was to discover the amount of vacancy in different sections of the city, the extent of vacancy in single and multiple family dwellings, and whether more or less vacancy existed in buildings designed jointly for business and dwelling occupancy. The sampling method was used, every sixth census enumeration district being canvassed. About 24,000 of the 140,000 families in the city were included in the study. A card was filled out for each dwelling place; hence the number of cards equalled the total number of places which either were or could have been occupied by a family. Thus for a three-family house three cards would have been turned in with the same address, each card recording the status of a single flat in the building.

The cards were kept separate by enumeration districts. After they had been edited and the number of cards from each district recorded on a master sheet, the cards were ready to sort. They were first distributed in five piles according to the number of dwelling places in the building. Each pile was then sorted into residential or combination residential and business. Each of these ten piles was sorted according to whether the dwelling place was occupied or vacant. The number of cards in each of the twenty piles was then counted and the results entered in the proper row of Table 12.

FIGURE 13
COLLECTION CARD USED IN RESIDENTIAL VACANCY INVESTIGATION
IN BUFFALO, NEW YORK

Serial No.

Address

Ward Tract Enumeration District

No. of Dwelling Places in Building:

One	Two	Three
Four	Over Four (give number)	
Occupied	Vacant	
Residential	Combination	

Agent

The cards were then collected into a single pack, shuffled, and turned over to another tabulator to be sorted, counted, and entered independently on a duplicate of Table 12. The cards were then held in distributed form until a third person checked the two records together and checked the total of the row with the number of dwelling places in that district as shown on the master sheet. If the two records agreed, and the totals checked with the original count, they were considered to be correct. If not, the cards were recounted until the discrepancy came to light. A similar procedure was followed in each enumeration district. As indicated in Table 12, the results were sub-totaled by tracts, if there was more than one enumeration district in a tract, and in every case by wards. This plan made it possible to use whichever of the geographic subdivisions was desired in subsequent work.

Tally Sheet

The use of a tally sheet is the reverse of sorting-counting in that the schedule cards or sheets are not separated into piles according to the various classifications. Instead a blank form, or several of them in a complex investigation, is made up to conform to the classifications of the data. The information is then scored on the blank form as it is read from the collection form. One person should read from the

TABLE 12
PRELIMINARY TABLE USED IN RECORDING RESIDENTIAL VACANCY INFORMATION IN BUFFALO, NEW YORK

		DWELLING PLACES IN BUILDINGS ACCOMMODATING															
WARD No.	TRACT No.	ENUMER- ATION Dis- trict No.	ONE FAMILY			TWO FAMILIES			THREE FAMILIES			FOUR FAMILIES			OVER FOUR FAMILIES		
			Resi- dential		Combin- ation	Resi- dential		Combin- ation	Resi- dential		Combin- ation	Resi- dential		Combin- ation	Resi- dential		Combin- ation
			O*	V*	O	V	O	V	O	V	O	O	V	O	O	V	O
		11	169	1	7	3	104	0	0	0	0	0	0	0	0	0	0
	2	8	63	2	15	1	160	0	12	0	0	0	0	0	0	0	0
	3	6	28	0	4	1	30	0	2	0	0	0	0	0	0	0	0
	4	2	93	1	6	0	72	0	2	0	0	0	0	0	0	0	0
	5																
1			353	4	32	5	366	0	16	0	0	36	0	9	0	4	0
		18	119	0	1	0	40	0	4	0	0	0	0	0	0	0	0
	6	20	154	0	0	1	142	0	0	0	0	6	0	0	0	0	0
			273	0	1	1	182	0	4	0	0	6	0	3	0	0	0
		13	43	0	0	0	78	0	0	0	0	3	0	0	0	0	0
	8	15	154	0	4	0	82	0	2	0	0	3	0	0	0	0	0
			197	0	4	0	160	0	2	0	0	6	0	0	0	0	0
	2		470	0	5	1	342	0	6	0	0	12	0	3	0	0	0

* O = occupied, V = vacant.

schedules while another person records on the tally sheet. There is an advantage in having several persons record the information simultaneously on separate sheets in order to secure one or more checks from the same reading. This, however, provides no check on the accuracy of the reading, so that perhaps the safest procedure is two independent readings and recordings. The weakness of the method lies in the fact that an error can be corrected only by rereading all the schedules. A device which partially overcomes this weakness is to divide the schedules into piles and then subdivide the tally sheet into corresponding parts. An error can then be localized in one of the piles and the rereading confined to that one.

If this method were to be used in tabulating the information of the vacancy survey in Buffalo, the tally sheet would have the form shown in Figure 14. One person would read off first the number of dwelling places in the building, then whether the building was residential or combined business and residential and then whether occupied or vacant. The person doing the tallying would locate the proper block as the information was read and would register one stroke for each dwelling place. The subtotals by enumeration districts, tracts, and wards would be recorded as indicated on the tally sheet.

If there is too much cross-classification of the data, obviously this method becomes cumbersome. In that case it is probably better to abandon the tally sheet and use sorting-counting or the work-sheet method explained in the next section. The tallying process may be simplified, however, by sorting the schedules first into their major classifications and then tallying by subgroups.¹ In Figure 14 the cards would be separated according to enumeration districts before the tallying was done.

The tally-sheet method is often the most desirable to use in taking information from a published source. For example, if one wished to record the number of industries in an area according to number of employees, the best method would be to make up a tally sheet with a classification by number of employees and tally the industries as they were read from the *Census of Manufactures*.

The Work Sheet

The purpose of a work sheet is to bring the information together in more convenient form than the schedules, so that it will be ready for further tabulation and analysis. After having been edited, the

answers recorded on all the schedules are transferred to a single sheet or several sheets, depending on the size of the investigation. The headings of these sheets will correspond to the questions on the collection schedules. Thus any tabulation which can be obtained from the original collection forms can be taken equally well from the work sheet.

In chapter VI, page 105, a questionnaire was presented concerning the college market for radios. Figure 15 is a proposed work sheet for recording this information. One row of the work sheet is used for each questionnaire, thus the identity of the information is preserved.

At least one sorting of the schedules can be made prior to recording any of the information on work sheets. In this case, the first obvious question on which to sort is whether the student now owns a radio. Second, from the non-owners, those who have not expressed the intention of purchasing a radio during 1936 may be eliminated entirely. There is no information that needs to be tabulated regarding this latter group, except a count of the total number of such cases. The complete form of work sheet shown in Figure 15 is needed only for the schedules of present radio owners. The non-owners who expect to buy can be recorded on a much shorter form that includes only the last three sections which deal with expected purchases in 1936. This procedure not only saves time in recording and space on the work sheets, but simplifies the task of classifying the information that will later be taken from these work sheets.

Each section on the work sheet represents one question on the schedule and must include enough separate columns to accommodate every possible reply that is expected to that question. In transferring the information from the schedule, a check mark or other equivalent symbol is made in one column of each section. Thus the number of cases in any column can be totaled easily by counting the check marks in that column. The sum of the column subtotals in every section should all give the same result, which should be the total number of schedules being tabulated.

In planning the headings of the sheet, it must be anticipated that for practically every question there are likely to be unexpected replies or unusual cases requiring notes of explanation, as well as instances where no answer appears. It is necessary therefore to provide extra columns in nearly every section, such as those in Figure 15 marked "Notes," "Other" and "Don't Know." By checking "Don't Know,"

a count of such cases can be included in the check total for that section. All written notes are confined to "Notes" and "Other," and do not interfere with the count of check marks.

The simple process of totaling single columns becomes more complicated when cross-relationships are wanted. For example, if the make of the present radio were to be related to the make of the radio the student expected to buy during 1936, a two-way table would be prepared and the tallying method used. Only those who answered "yes" to the question, "Will you purchase a radio before 1936?" would be included. The completed work for 100 cases might appear as shown in Table 13. The facts could be read in the form shown, or as a final table by substituting figures for the tally marks.

TABLE 13
PROPOSED TALLY SHEET USING PART OF INFORMATION RECORDED
IN WORK SHEET, FIGURE 15, HYPOTHETICAL DATA SHOWING
RELATION BETWEEN MAKE OF RADIO COLLEGE STUDENTS
OWN AND MAKE THEY EXPECT TO PURCHASE

MAKE OF RADIO OWNED	MAKE OF RADIO EXPECTS TO PURCHASE										Total
	A	B	C	D	E	F	G	H	Other	No Informa- tion	
A	III	I			II		I	I	I	II	11
B		II	I	I			I		I	III	9
C	I	II		I		II		I	I		8
D	I			II		I				I	5
E		I	I	I	III		I		I		8
F		I		I		III		I			6
G	I	II			I		III		I	I	10
H			I		II			II	I		6
Other	II	I		III	I		I	I	III	II	15
Home made	I	I			I	I		I		II	7
No informa- tion				II		I	I		I	III III	15
Total	9	11	3	11	10	8	9	7	11	21	100

After this has been done, it can readily be seen that the collected information is no longer in a preliminary state. The table represents a selection and combination of certain parts of the original data and can be read as follows:

a) Fifteen per cent did not state the make of radio owned and 21 per cent did not state the make of radio they would purchase. It is difficult to draw any conclusions from the remainder of the table with so much information missing.

b) Radio "C" seemingly is in disrepute. None of the present owners would repurchase it and only three would turn to it from other makes.

c) While there is considerable evidence of shifts in consumer preference, the radio owned at present has some advantage over competing products except in the case of "C."

d) At least five of the students who built their present sets would not do so again.

Other combinations of data can be made from the information on the work sheet, using tally forms similar to that shown in Table 13. The purpose for which the information was gathered will determine what forms to use. Note, however, that the work sheet itself is in no sense a final form for the data but merely an intermediate device. It is seldom published even when a complete record of the statistical analysis is included along with the report of an investigation.

Mechanical Tabulation

When a large number of schedules is to be analyzed the task of tabulation becomes enormous. Likewise when a great amount of cross-tabulation is necessary, even though the number of cases is not so large, the task of preparing tables is likely to become the "bottle-neck" of the investigation. Under either of these circumstances the present practice is to abandon hand tabulation in favor of the use of machinery designed for the purpose. At no other point is the statistician so much favored by the developments of the machine age as in tabulation. Equipment is available to perform quickly and accurately the steps of sorting, counting, cross-tabulating, and recording in columnar form.¹

These advantages have led a great many business concerns to install mechanical systems for the maintenance of records of current operations.¹ The variety of uses made of the "punch card" system for this purpose is illustrated by the following examples: bad debt losses of members of a retail credit association; broker's record of security dealings with customers; merchandise control of a mail-order house; records of premium payments of a life insurance company; service record

of employees of a shipbuilding company; deliveries to individual stores from the central warehouse of a chain grocery company; stock control in the warehouse of a chain grocery company.

Principles.—The basic principle of machine tabulation is that a hole punched in a card represents by its horizontal and vertical position a certain statistical fact. It becomes a permanent record that can be used in tabulation at any time by running the card through a machine.

The first machine developed for this purpose was the "sorter." This machine will sort a pack of punched cards into numbered compartments according to any one set of information. Further refinement has led to an attachment which will count the number of cards going into each compartment; sorting-counting has thus been reduced to a single operation.

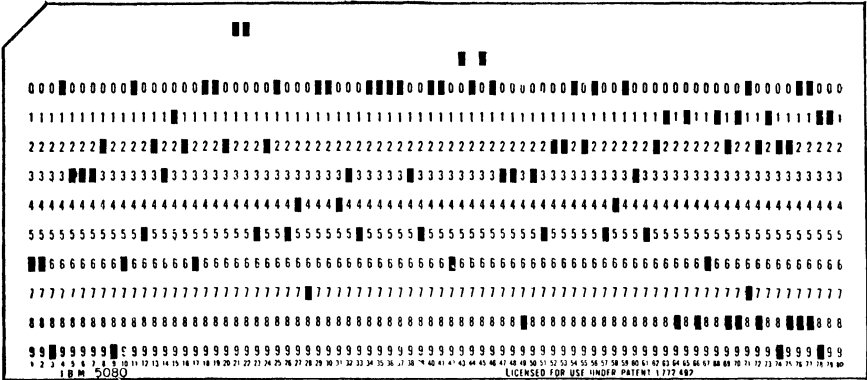
The next step was the invention of the "tabulator," a machine that operates at a more complex level. After the cards have been sorted, the tabulator can add the amounts recorded on each and furnish a printed record of the total. For example, if cards were punched showing the weekly wage rates of a firm's employees, each card representing one employee, the tabulating machine could be set so that it would give a printed record of the number employed at each wage rate, the total earnings of each group, the total number of employees, and the total weekly payroll with a single running of the cards.

Steps in the Process.—Probably the only way to understand fully what the machines can do is to see them in operation.² It will be worthwhile, however, to present in some detail those parts of the process which receive the least attention in a practical demonstration: (1) preparation of a code; (2) transfer of collected information to a code sheet; (3) punching the cards; (4) sorting-counting; (5) tabulation of numerical information; (6) cross-tabulation; (7) recording in tables. Some types of data will not require the use of all of the steps, but steps 1, 3, 4, and 7 will always be necessary.

The code: The preparation of the code to be used in transferring information from long-hand forms to a set of holes punched in cards such as those shown in Figure 16 requires considerable skill. Each card

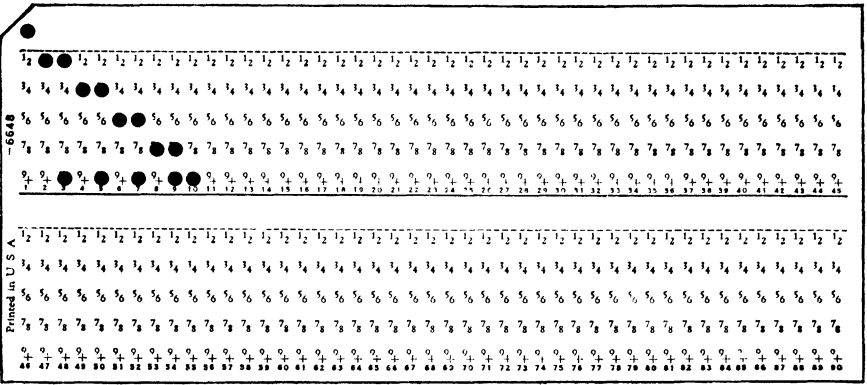
² These machines are manufactured by the Electric Accounting Machine Division of the International Business Machines Corporation and the Tabulating Machine Division of Remington Rand Inc. The authors have found local representatives of both companies willing to demonstrate the use of the equipment either to individuals or to groups of students. Seeing the machines at work is vastly superior as a teaching device to mere description of their operation even though aided by pictures.

FIGURE 16
PUNCHED CARDS FOR MECHANICAL TABULATION
A. 80-column card (punched as coded in Figure 19)*



* Reproduced through the courtesy of the Electric Accounting Division of the International Business Machines Corporation.

B. 90-column card * †



* The 90-column card is made possible by dividing a 45-column card horizontally, allowing six spaces in each half column. The machines can be set to use either the upper or lower half of the card. The six spaces are used for 11 numbers by means of a combination punch, e.g., in column 1 the space above the line is punched for zero; if the 1s space alone is punched as in column 2 it represents 1; 1s and 9+ as in column 3 represents 2; etc.

† Reproduced through the courtesy of the Tabulating Machine Division of Remington Rand Inc.

usually represents one schedule or other individual record. Each question on the schedule is represented by one or more vertical columns on the card, while the numbers 0-9 in the column signify the various possible answers to that question.

The actual work of preparing a code can be visualized from the example shown in Figures 17 and 18. Figure 17 is a reproduction of

page 1 of a schedule used in collecting housing information in Buffalo, New York, in 1930. Figure 18 is a reproduction of the instructions for coding the answers to questions A-1, A-2 and A-3 (*a*), (*b*), (*c*), and (*d*). Columns 1-3 make a direct transfer of the serial numbers of the schedules to the card. Column 4 shows the simplest type of transfer of non-numerical information to the card. The note attached to the code for column 4 indicates that the information in question A-2 was not coded. This was omitted because only a few of the houses in the study showed any variation in construction material. Column 5 involves the same transfer of non-numerical information as column 4, although not so easy to record because each number in the column stands for a combined occurrence of cellar and attic. Column 6 is a simple transfer of numerical information. Column 7 is a combination transfer like column 5. Columns 8 and 9 use a somewhat complicated plan for recording the other rooms in a house. For example, four in column 8 and seven in column 9 means that the house contained a dining-room, an entrance hall, an inclosed porch, an open porch, and one other room. In the same way other numbers in the two columns indicate various combinations of rooms in a house. Column 10 gives a summary of the detailed information recorded in columns 7, 8, and 9. The note appended to the code for column 10 is necessary as a guide for the coder.

In columns 7 and 10 the symbols X and B appear. The machines will sort on 12 items in a column since two extra punches can be made in the upper margin; therefore there is provision for two extra items in a column if necessary. This was done in column 7 to allow for two additional combinations of bedrooms and baths. Likewise in column 10 the largest house in the study contained 11 rooms and a few schedules were indefinite on the total number of rooms.

It is not sufficient to prepare a code that will provide for all the information on the schedule, but the code should be so arranged that the desired tabulations can be taken from the machines with a minimum of sorting and cross-tabulating. In short, the person preparing the code must be familiar with all of the machine operations as well as with the method of subsequent analysis.

For example, a code may be required that will designate each of the 48 states and the District of Columbia. It is obvious that two columns are needed in order to include so many items. Experience shows that to use such a code as Alabama-00, Arizona-01, Arkan-

FIGURE 17

THE PRESIDENT'S CONFERENCE ON HOME BUILDING AND HOME OWNERSHIP

Form One (Partial Reproduction)

Case No.

Information from Home Purchasing Families

A. Description of house

1. Type of construction

Single house

Income house

Two-family house

Other (specify)

2. Material

Superstructure:

Cellar:

Roof:

Frame

Stone

Shingle

Brick veneer

Tile

(a) Plain

Brick

Concrete block

(b) Treated

Tile stuccoed

Other (specify)

Fibre shingle

Other (specify)

Slate

Other (specify)

3. Rooms and space included

(a) Is there a cellar Attic: Finished Unfinished

(b) How many floors including cellar and attic

(c) Is there a separate dining room dining nook or alcove
kitchen pantry laundry room entrance
hall

(d) Number of bedrooms bath rooms lavatories
glass enclosed porches screened porches open
porches other rooms (including living room, den, library,
play room, sewing room, etc.)

(e) Dimensions of house: Front Depth

(f) Size of lot occupied by house: Front Depth

(g) Value of lot without house:

At time of purchase

Present assessment

(b) Corner lot

4. Garage

(a) Size: One-car Two-car Three-car

(b) Construction:

Frame

Concrete block

Brick

Other (specify)

FIGURE 18
INSTRUCTIONS FOR CODING DATA ON HOME BUILDING AND HOME OWNERSHIP
IN BUFFALO, NEW YORK, 1930
(Reproduction of columns 1-10)

Column 1-3. *Serial number of schedule*

001 = 1
 526 = 526
 999 = 999

Column 4. *Type of house*

0—Single house 2—Income house
 1—Two-family house 3—Three-family house

(NOTE: Although material about the type of construction of the house will not be coded, the coder is instructed to keep a list of the schedule number of every house that is not of *frame* construction and to note the exact type of construction of these exceptions.)

Column 5. *Cellar and attic*

0—No cellar and unfinished attic 3—Cellar and finished attic
 1—No cellar and finished attic 4—Cellar and no attic
 2—Cellar and unfinished attic 5—No cellar and no attic

Column 6. *Number of floors in house*

1—One floor 3—Three floors
 2—Two floors etc.

Column 7. *Number of bedrooms and bathrooms*

1—1 bedroom and 1 bath	7—1 bedroom and 2 baths
2—2 bedrooms and 1 bath	8—2 bedrooms and 2 baths
3—3 bedrooms and 1 bath	9—3 bedrooms and 2 baths
4—4 bedrooms and 1 bath	0—4 bedrooms and 2 baths
5—5 bedrooms and 1 bath	X—5 bedrooms and 2 baths
6—6 bedrooms and 1 bath	B—6 bedrooms and 2 baths

Column 8-9. *Other rooms in house*

See Appendix 1 for code number to be used.

Column 10. *Total number of rooms in house*

1—1 room	0—10 rooms
2—2 rooms	X—11 rooms
3—3 rooms, etc.	B—Unknown

Rooms to be counted in computing total for Column 10

Include:

Exclude:

Dining-room	Other rooms	Dining nook
Kitchen	Living-room	Pantry
Entrance hall	Den	Laundry room
Bedroom	Library	Bathroom
Inclosed porches	Playroom	Lavatories
	Sewing room	Open and screen porch

sas-02, etc., is not efficient. It is better to use one column for geographical subdivisions and the second for the states in each as follows:

Maine	00	Connecticut	05
New Hampshire	01	New York	10
Vermont	02	New Jersey	11
Massachusetts	03	Pennsylvania	12
Rhode Island	04	etc.	

Then, if only the major subdivisions are needed in a table, it will be necessary to sort on only the tens column instead of both.

Other precautions can be taken to reduce to a minimum errors in transcribing data to the code sheet as well as in punching the cards. Whenever possible the number used in the code should correspond to the number written on the schedule. In Figures 18 and 19 it will be observed that this has been done for column 6, but not for column 4. In the latter case it was considered preferable to take off the answers in the same order in which they appeared on the schedule, but "income house" could have been placed first in preparing the schedule and coded as 0, one-family as 1, two-family as 2, etc.

For the same reasons that determine the inclusion of extra columns in a work sheet, it is desirable to make allowance in the code for "no answer" or special cases which may require listing by hand. As far as possible the same number should be used in each column for such answers, the 11th and 12th positions being used for this purpose. They are called "X" and "B," or "X" and "Y."

The code sheet: The complete code sheet reproduced as Figure 19 was used as an intermediate step between the transfer of the information on housing from the collection schedule (partially reproduced in Figure 17) to the punched cards. The sheet is divided into fields and each column is labeled to conform to the descriptions in the code (partially reproduced in Figure 18). The entries in each row of the body of the sheet are the code numbers which stand for the information contained in the schedule whose serial number appears at the left. Schedule No. 669 is recorded on Figure 19 to illustrate the procedure. This information is also reproduced on the 80-column card of Figure 16-A.

The construction of the code sheet greatly facilitates the punching

of the cards, but is not an absolutely necessary step in the process. The transfer of the information from written to numerical form can be done on the margin of the collection schedule. More time will be spent by the punch operator in taking the information from the margin of the schedule than from the code sheet. On the other hand the transfer from schedule to code sheet will require more time than coding on the margin of the schedule. Which method to use must be determined in each case in the light of the particular circumstances involved. The three primary factors to consider are time, expense, and accuracy.

For continuous recording of operations by business concerns the better plan usually is to arrange the original record so that the coding can be done on it directly without the use of a code sheet. The punch operator becomes so familiar with the code after a few months that the information can be punched directly from the original records without further reference to the code instructions.

Punching the cards: The punching machine is just a simple device with 12 numbered keys fixed over a movable carriage containing the card. As a hole is punched in a column according to the coded information the card advances automatically to the next column in position for punching. These punches may be operated by hand or by electricity.

Sorting-counting: This operation is the fastest part of the process. The machine sorts and counts several hundred cards a minute. In one type of machine the card itself acts as an insulator breaking an electric circuit. When the brush carrying the current comes to a hole in the card, the electric circuit is completed. The completed circuit in turn opens the guiding device to a compartment whose number corresponds to that of the hole in the card, and the card drops into the compartment as it passes along on a conveyor. The other type of machine performs the same process except that the cards are picked out by pins that drop into the holes, instead of by a completed electric circuit. Both machines also have attachments that will count the number of cards falling into each compartment. Since the machine has only 12 pockets, it is obviously not possible to sort according to more than one column at a time. When it has been necessary to utilize two or more columns in a single code (as in Figure 19, columns 23 and 24) the cards must be run through separately for each column, if a complete sort is wanted. In these two columns the cost of the house is recorded in hundreds of dollars, the last two ciphers having been

FIGURE 19
Code Sheet
 Used in Transferring Information From Questionnaire, Figure 17
 (Code filled in for schedule No. 669)

HOUSE DESCRIPTION													DATES			ORIGINAL HOUSE COST					MORTGAGE PAYMENTS					BREAD WINNER		FAM		BREAD WINNER		FAM		DOUBLE PUNCH																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																
SERIAL NUMBER													TYPE			NO FLOORS					NO BEDROOMS					OTHER ROOMS					NO ROOMS		HOUSE FRONT		HOUSE DEPTH		LOT FRONT		LOT DEPTH		GARAGE		YEAR BUILT		1ST SLD BLT		LAST SLD-BUILT		ORIG SELL PRICE		TOT COST PRES OWN		DOWN PAYMENT		ORIG VALUE		1ST MTGE		ORIG VALUE		2ND MTGE		AMORTIZATION		FIRST MORTGAGE		SECOND MORTGAGE		1ST AND 2ND MORTGAGES		HOLD RE-REFIN		WRITEN REFIN		AGE		OCCUR WEEKS EMPLOY.		TOTAL EARN.		TOTAL EARN. DEPEND		AGE		OCCUP WEEKS EMPLOY.		TOTAL EARN		TOTAL EARN DEPEND		DOWN PAYMENT		EDUC VAC		AUTO HELP		FURNISH. EQUIP		CLOTHES		INS		SAVINGS		MOVIES THEATRES		BOOKS MAGAZINES		HOME IMPR		ENF SAV. INT		JOB SCHOOL		MOVING																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																															
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35</

dropped. Thus if a count is wanted by \$100 groups, the cards are first sorted on column 23, in \$1,000's, and each \$1,000 is then re-sorted and counted on column 24 in \$100 groups—\$2,000, \$2,100, \$2,200, etc

Tabulation and cross-tabulation: The "tabulator" consists of several banks of small adding machines which may be arranged electrically to record, total, and print almost any desired combination of data from the cards. The same thing can be accomplished by the use of the sorter alone, but only after multiple sorting, re-arranging, and computing. The difference can be demonstrated by further reference to Figure 19.

It was desired to prepare a table showing the average cash payment according to the total price to the present owner in \$1,000 intervals. Without the tabulator, it would be necessary to make a sort on column 23 and then to re-sort and count each pack separately according to columns 25 and 26. This would give results in the form of 12 frequency distributions each having possibly 100 class intervals from which the average cash payments could be derived. With the tabulator it was necessary to sort only on column 23.⁸ The tabulator was then set to count the items in each \$1,000 price group, to add the exact amounts recorded in columns 25 and 26 and to print the subtotals and grand totals as shown in Figure 20, columns 1, 2, and 3. The amounts of the first and second mortgages, columns 4 and 5, were also totaled during this same operation. Reading across the first row, the total cash payments made by 11 purchasers of houses costing between \$2,000 and \$3,000 amounted to \$4,700; the total value of the first mortgages of these 11 houses was \$17,900 and the total value of the second mortgages was \$4,800.

Recording results in tables: When the sorter alone is used a large number of blank forms must be prepared in advance for the recording of any possible cross-tabulation that may be needed, but with the tabulator this task becomes much lighter,

⁸ When used in conjunction with the tabulator, the principal function of the sorter is to arrange the cards in consecutive order according to some one classification which will become the stub of the resulting table. If this classification has required the use of two columns on the card, the right hand or *units* column should be sorted first. These ten packs are then piled together in order, with zeros at the bottom and are run through the sorter on the left hand or *tens* column. Since the cards at the bottom of the pile pass through the sorter first, the unit zeros are sorted first and fall at the bottom of each tens compartment, followed by the unit ones, twos, etc. When the tens packs are then piled together, with zeros at the bottom, the entire set of cards is in consecutive order, and is ready to tabulate. The tabulator can be set to separate them according to each unit, if desired. In the case illustrated, only the tens values were needed, so it was necessary to sort and tabulate only on column 23, disregarding column 24 entirely.

FIGURE 20
REPRODUCTION OF THE PRINTED RECORD FROM THE
TABULATING MACHINE WITH HEADINGS ADDED,
DATA FROM CODE SHEET, FIGURE 18

(1) COST OF HOUSE TO PRESENT PURCHASER (\$1,000)	(2) NUMBER OF CASES	(3) TOTAL OF CASH PAYMENT (\$100)	(4) TOTAL VALUE OF FIRST MORTGAGES (\$100)	(5) TOTAL VALUE OF SECOND MORTGAGES (\$100)
2	11	47	179	48
3	27	238	411	302
4	32	357	772	296
5	158	1186	4631	2748
6	134	1816	4574	2091
7	95	1998	3559	1497
8	62	1173	2685	1294
9	41	932	1999	842
0	23	629	1178	532
†	14	417	798	338
†	22	968	1292	526
	619*	9761*	22078*	10514*

† The information in this row was recorded in the B position of the cards. The tabulator did not print the B.
* Asterisks denoting totals are part of the machine record.

The first step is the preparation of the plan for the derivative table which will result from the tabulation. Then an outline form of the primary table is made. This is used by the machine operator in arranging the machine. When the cards are tabulated the results are printed by the machine in a form similar to Figure 20. This is the primary table. From the primary table the derivative table can be constructed.

TABLE 14
DERIVED TABLE FROM TABULATING MACHINE RECORD, FIGURE 20:
ARITHMETIC AVERAGE OF CASH PAYMENT AND ORIGINAL FACE VALUE OF
FIRST AND SECOND MORTGAGES FOR DIFFERENT PROPERTY COSTS.
RESTRICTED TO PROPERTIES PURCHASED IN 1922 OR AFTER.
619 BUFFALO FAMILIES

COST OF PROPERTY TO PRESENT OWNER	No. OF CASES	AVERAGE CASH PAYMENT	AVERAGE FACE VALUE FIRST MORTGAGE	AVERAGE FACE VALUE SECOND MORTGAGE
\$2,000- 2,999	11	\$ 430	\$1,630	\$ 440
3,000- 3,999	27	880	1,520	1,120
4,000- 4,999	32	1,120	2,410	920
5,000- 5,999	158	750	2,930	1,740
6,000- 6,999	134	1,360	3,410	1,560
7,000- 7,999	95	2,100	3,750	1,580
8,000- 8,999	62	1,890	4,330	2,090
9,000- 9,999	41	2,270	4,880	2,050
10,000-10,999	23	2,730	5,120	2,310
11,000-11,999	14	2,980	5,700	2,410
12,000-15,999	22	4,400	5,870	2,390
Total or average	619	1,580	3,570	1,700

In this case Table 14 was constructed by dividing the totals in each row of columns 3, 4, and 5 of Figure 20 by the number of houses in the row, column 2.

Summary of Mechanical Tabulation.—It is hardly to be expected that the whole process of mechanical tabulation will be clear merely from reading this description. Perhaps it will serve as a guide as to what to look for in a demonstration of the machines.

Mechanical tabulation is a great assistance in accounting and statistical work both internal and external. But it cannot be used to advantage in all cases. Certain types of work call for the mechanical process; others are not adapted to its use. The outstanding criterion is the size of the investigation. If either a large number of cases or a large amount of cross-information is involved, the mechanical process should be used.

PROBLEMS

1. What routine does an editor follow in searching for irregularities in schedules?
2. What is meant by re-editing? When should the process be employed?
3. What would you include in the qualifications of an editor?
4. The following is a card from the investigation illustrated in Figure 13. This card was returned to the collecting agent by the editor. What do you think the editor found wrong and what did he want the agent to do?

COLLECTION CARD USED IN RESIDENTIAL VACANCY INVESTIGATION IN BUFFALO, N. Y.

Serial No.

Address 526 So. Elm St. front and rear houses

Ward Tract Enumeration District

No. of Dwelling Places in Building

One.....	Two.....	Three.....
Four...4.....	Over Four (give number).....	
Occupied <u>2</u>	Vacant <u>1</u>	
Residential <u>X</u>	Combination <u>X</u>	

Agent A. S. Agnew

5. Describe in detail an investigation in which you would use *sorting-counting* at the preliminary tabulation stage.
6. Describe in detail an investigation in which you would use a *tally sheet* at the preliminary tabulation stage.
7. Describe in detail an investigation in which you would use a *work sheet* at the preliminary tabulation stage.
8. What is the principle of mechanical tabulation?
9. What are the advantages of mechanical tabulation?
10. Describe an investigation in which mechanical tabulation should *not* be used.
11. The following is an approximate reproduction of the invoice form used by a firm having 63 salesmen, 4,000 customers, and listing in its catalogue 13,200 commodities classified in 12 departments.

JONES SMITH INC.

HEAVY HARDWARE

"Serving the United States"

Dept.

Date

Salesman

Invoice No.

Territory

SOLD TO

Quantity	Commodity	Unit Price	Amount	

Prepare a code for transferring all of the information on this invoice to 45-column mechanical tabulation cards. The code should be prepared so that information could be taken from the cards concerning any of the following either separately or in combinations: sales from day to day, sales by departments, the record of any one individual invoice, the amount of goods sold over a period to any individual customer, the distribution of sales by states, cities, or territories, the sales made during a month by each salesman, the quantity of each commodity sold during a period, the amount of sales during a period.

REFERENCES

BAILEY, WILLIAM B., and CUMMINGS, JOHN, *Statistics*. Chicago: A. C. McClurg & Co., 1917.

Chapter IV contains a complete and lucid statement of the purpose of editing schedules and the duties of an editor.

BROWN, LYNDON O., *Market Research and Analysis*. New York: The Ronald Press Co., 1937.

An excellent statement on editing and methods of testing a sample appears in chapter 12 and rules for preliminary tabulation are given in chapter 13.

EIGELBERNER, J., *The Investigation of Business Problems*. Chicago and New York: A. W. Shaw Co., 1926.

Chapter XIII explains the function of the editor as official critic of the quality of collected information.

SCHLUTER, WILLIAM C., *How To Do Research Work*. New York: Prentice-Hall, Inc., 1929.

A concise statement of the editing process appears in chapter XI.

Standards of Research. Des Moines, Iowa: Meredith Publishing Co., 1929.

An excellent statement of the functions and qualifications of an editor appears on pages 29–32. Instructions for mechanical tabulation are given on pages 32–36 and a standard code on pages 69–75.

CHAPTER VIII

TABULATION

DEFINITIONS

THE TABULATION of statistical data is the orderly arrangement of concrete numerical information in vertical columns and horizontal rows. This definition excludes lists of facts which are not numerical and mathematical tables which deal with abstract numbers. It contains three separate concepts: (1) numerical information regarding actual items, events, values, or relationships; (2) a definite order of arrangement for this information; (3) the preparation of forms with rows and columns in which the numerical data may be recorded according to their orders of arrangement.

The method of collecting concrete numerical information has been discussed in chapters IV and VI. It has already been indicated in these chapters that the final form in which the collected data are to be arranged must be anticipated to some extent in planning the collection, and also in the stages of preliminary tabulation as explained in chapter VII. However, the thorough analysis of the various orders of arrangement has been reserved for this chapter. The latter part of the chapter is concerned with the third factor in tabulation, the principles and practices in preparing tabular forms for the recording of classified statistical data.

Two kinds of statistical information may be presented in tabular form: (1) several sets of more or less heterogeneous information, and (2) data representing a definite universe expressed in a common unit. In the first kind of table the several sets of information are not expressed in the same unit, but they are arranged according to a single common characteristic, such as the dates of successive observations. Grouping such data in tabular form is a space-saving device very often used and is a legitimate statistical technique provided that the different sets of information bear some relation to each other. The other kind of table contains homogeneous data employing a common unit and arranged according to one or more definite orders of classification. The ensuing discussion of the elements and orders of classification deals with the second kind of table.)

Classification

| *Classification is the arrangement of a set of observations or computed figures according to a previously determined plan. This arrangement may involve the separation of a whole into parts or the listing of related sets of information.*¹

Elements of Classification.—Each plan of classification involves several indispensable elements:

1. The data have been collected and arranged for some definite statistical purpose.
2. The enumeration or collection has involved one unit, that is, the items counted were all defined in the same way.
3. These similarly defined units each possess one or more of the same variable characteristics so that all of them can be classified according to each of these characteristics.
4. For each classification that involves the separation of a whole (that is, a total of identically defined units) into its parts, the classes must be mutually exclusive.

The foregoing statements all refer to data in the original form in which they have been collected, that is, to primary data. The distinction between primary and derived tables will be mentioned later, but for the present the elements of classification and the various orders of classification are discussed in the simplest terms, as applied to primary data.

Purpose: The purpose is so closely connected with the other elements of classification that it requires little elaboration. For example, we wish to know how many students in a freshman class are men and how many are women. "Freshman students," therefore, is selected automatically as the unit to be counted, and the variable characteristic according to which the units will be classified is "sex." Since this variable can have only two classes, male and female, a very simple classification will result:

¹ This definition differs materially from that given in Edmund E. Day, *Statistical Analysis* (New York: The Macmillan Company, 1925), pp. 36 and 42. Day distinguishes *classification*—the separation of a larger group "population" or "universe" into smaller groups or classes on the basis of a specified criterion or characteristic feature—from *seriation*—the arrangement of an orderly succession of items relating differences in one variable to differences in another. This distinction, although theoretically a fundamental one, is disregarded in the discussion of the principles and methods of tabulation in this chapter, because classification and seriation both follow the same rules with regard to *arrangement*. The distinction made by Day is referred to at the end of chapter XIII, in introducing graphic methods that involve two variables.

Male	125
Female	98
Total freshmen	223

This elementary illustration shows why the final objective must be kept clearly in mind at the initial stages of the investigation. If no provision were made for recording *male* or *female* on the student's registration card, there would be no reliable way of counting the number of men and women, since such names as "LaVerne" or "Marion" may be of either sex.

Units: In any table of primary data, the unit is whatever is being counted. One should be able to select any single figure from a table, for example, the first figure from Table 1, chapter I, and ask, "804,350 what?" The answer which can be read from the title of the table is always in terms of the unit forming the basis of that particular tabulation, and *it will be equally applicable to every figure in the table.* In this case the unit is *passenger automobiles sold*. Note that, although one might also say 804,350 *Chevrolets* were sold in 1937, these further limitations do not apply to *all* of the figures in the table. They are therefore not part of the definition of the unit of the table. *Chevrolet* is one class of one of the variable characteristics "make of automobile" and 1937 is one class of another variable characteristic "model year."

Variable characteristics: Collected data may possess one or more variable characteristics according to the degree of detail necessitated by the purpose of the investigation. In the foregoing illustration, the variable characteristic "make of automobile" was subdivided into the classes, Chevrolet, Ford, and Plymouth, and the variable characteristic "model year" also had three classes, 1937, 1938, and 1939. In the previous illustration the freshmen who were classified by sex might have been differentiated according to a second characteristic, "university division." If a similar count were made in several universities, for several successive years, the complete data would contain four variable characteristics: university, division, sex, and year of entrance. Each student counted could be distinguished according to all four of the characteristics; for instance, student No. 1 might be a man, at University C, in Business Administration, entering in 1940.

Mutually exclusive classes: In the total count, therefore, each student would fall into *one*, and only one, of the possible categories under each of the four variable characteristics, that is, in one class of each classi-

fiction. The class "entered in 1940" could not include any of those counted as "entered in 1941"; none of the freshmen entering University A could also be counted as entering University B, etc. Likewise in Table 1, chapter I, each one of the 4,613,424 passenger cars sold during the 3 years is included in only one of the individual figures (known as "cells") of the table, except that the figures of the last row are subtotals, each one including all items in the column above it.

It will be explained later that classification according to two or more characteristics may be *crossed* in a single table, as in the table of automobile sales. However, there will be no overlapping of the several classes in any one classification unless two or more variable characteristics become confused in listing the classes. If this is allowed to occur, such classifications as the following may result:

HOMICIDES IN THE CITY OF NEW YORK, 1926

Manhattan	218	Shooting	213
Brooklyn	82	Assault	53
Bronx	19	Stabbing	54
Queens	22	Gas	7
Richmond	3	Infanticide	7
Whole city	344	Poison	0
By Negroes	41	Accidental	10
By husbands	4	By police	19
By wives	3	Suicide	10

In this table one must assume that the figure in the sixth row (whole city) is the total, since it is the sum of the figures for the five boroughs. These six rows alone comprise a correct tabulation of the whole number of homicides divided into five mutually exclusive classes according to the single characteristic, place of occurrence. The remainder of the table lists homicides according to at least three additional characteristics: race of the person committing the crime, relationship to the victim, and cause of death. If all of the classes of these several classifications were given so that each classification included all of the 344 cases, the table would not be incorrect, although it would afford no cross-classification. As it stands, all the classes appear to be parts of a single classification, but there is overlapping between them. For instance, some of the 41 homicides by Negroes probably were committed in Manhattan, some of the Negroes may have been husbands of the victims, and some of the stabbing cases may have also been committed by Negroes. A single homicide could have answered to every characteristic that has been suggested; hence the classes are not mutually exclusive.

Orders of Classification, Time, Space and Attribute.—There are three main types of characteristics with respect to which data may be classified: (1) variation in time, (2) variation in space, and (3) variation of an attribute. In a previous example freshman students were classified according to (1) time (year of entrance), (2) space (location of the university), and (3) the two attributes, sex, and university division in which each one was registered. Thus all three orders of classification are represented in this example. Another case of the three orders in a single set of data can be found in the tabulation of yearly car loadings in the United States, by "months," by "regions," and by "size of railroad." On the other hand many tables will contain only one or two orders of classification, or there may be two or three variable attributes and no variation in either time or space.

In all of the examples that have been mentioned, the number of units in each class was obtained by counting objects, that is, freshmen, automobiles sold, loaded cars, and even the cases of homicide, according to their several variable characteristics. In many tabulations, however, counted objects or persons are replaced by prices or rates. The data may be expressed in units of dollars and cents but each figure is actually a ratio meaning "number of dollars paid per bushel" or "per ton," etc. Such data may be classified in the same way as countable units, according to mutually exclusive variable characteristics, which may relate to changes in either the numerator or the denominator of the ratio.

Table 15 contains three such classifications of wheat prices, each of which illustrates one of the orders of classification. Table 15-A shows changes in the unit "Average Weekly Cash Price of No. 2 Hard Winter Wheat at Chicago" classified according to *time* of occurrence. The four weeks quoted constitute four classes in this time classification. Other characteristics are held constant by the definition of the unit, i.e., all prices are taken from the same market, for cash sales, for the same grade of wheat. There is no overlap in the classes of the classification.

Table 15-B shows changes when the unit "Average Cash Price of No. 2 Hard Winter Wheat for the Week of October 3-8, 1938" is classified according to the variable characteristic *place* of occurrence. The three cities named as markets are three mutually exclusive classes in the spatial classification. Other characteristics are held constant by

the definition of the unit, i.e., all prices are taken for the same time period, for cash sales, for the same grade of wheat.

Table 15-C shows changes in the unit "Average Cash Price of Spring Wheat at Minneapolis for the Week of October 3-8, 1938." The five grades of wheat in the attribute classification are clearly defined and non-overlapping in accordance with the definitions established by the United States Grain Standards Act. Other characteristics of the unit are held constant by the definition, i.e., all prices are taken for the same time period, for cash sales, in the same market.

TABLE 15
THREE TYPES OF CLASSIFICATION *

A. TIME

AVERAGE WEEKLY CASH PRICES OF NO. 2 HARD WINTER WHEAT AT CHICAGO
FOR FOUR WEEKS OF OCTOBER, 1938

WEEK	AVERAGE PRICE PER BU.
Oct. 3- 8	\$.669
Oct. 10-15	.672
Oct. 17-22	.674
Oct. 24-29	.680

B. SPACE

AVERAGE CASH PRICES OF NO. 2 HARD WINTER WHEAT IN THREE MARKETS FOR THE
WEEK OF OCTOBER 3-8, 1938

MARKET	AVERAGE PRICE PER BU.
Chicago	\$.669
Kansas City	.638
St. Louis	.678

C. ATTRIBUTE

AVERAGE CASH PRICES OF FIVE GRADES OF SPRING WHEAT AT MINNEAPOLIS FOR THE
WEEK OF OCTOBER 3-8, 1938

GRADE	AVERAGE PRICE PER BU.
Dark Northern Spring Heavy.. No. 1	\$.738
Dark Northern Spring..... No. 1	.733
No. 2	.701
Northern Spring	No. 1
No. 2	.640
Hard Amber Durum..... No. 2	.651

* *Crops and Markets*, United States Department of Agriculture, Vol. XV, No. 11 (November, 1938), p. 254.

The three types of classification can be easily distinguished by noting that in Table 15-A all observations refer to a single place and invariant attributes, time being variable; in Table 15-B all observations refer to a single time interval and invariant attributes, location or space being variable; and in Table 15-C all observations refer to a single time period, a single place, and attributes that are invariant except the

attribute of grade of wheat. Greater care is necessary in dealing with variable attributes than with variable time or place because a universe can have many different attributes. For example, classifications of prices of spring wheat could also be made according to percentage of dockage, kind of contract of sale, or type of purchaser.

Attribute classifications are sometimes divided into qualitative and quantitative. Table 15-C is an example of a division according to the qualitative attribute, grade of wheat. Attribute classifications are quantitative when the attribute is expressed numerically, as in size or price groups. Such classifications take the form of frequency distributions, the discussion of which is deferred to chapter XV.

TYPES OF TABLES

Statistical tables can be divided into two categories—primary and derivative.

Primary Tables

A primary table is a full presentation of the collected data in the original units. An investigation of any complexity may require several primary tables. Such complete tables serve as a basis from which the statistician selects certain related sets of data that may be presented in various ways, depending on the purpose in view.

If the original data are to be published for general use without knowledge of what the uses will be or which relationships will be considered most important, primary tables may be given in full. Due to the expense of publication such tables are not commonly found in print, but certain parts of the original data such as a group of subtotals comprising a grand total may be published in order to bring the important data together in compact form.

Derived Tables

A derived table is one that presents the results of some analysis of the original data, such as percentage distributions, per cents of increase or decrease, values per capita, index numbers, or coefficients. These are constructed from the original data by the application of statistical methods and may be published either alone or accompanied by a part or all of the data upon which they depend.

The chief requirement of a derived table is that it should present one unified set of relationships. An attempt to set forth too many ideas in one derived table usually results in confusion. The preferable method is to use several short clear tables each of which has one definite purpose. Thus one primary table frequently becomes the source for many derived tables.

Parts A and B of Table 16 show two entirely different sets of information, but both were drawn from the same primary table which gave the distribution of explosives workers of each grade of skill according to average hourly earnings.

TABLE 16
TWO DERIVED TABLES FROM ONE PRIMARY TABLE*

A

PERCENTAGE DISTRIBUTION OF EXPLOSIVES WORKERS, BY AVERAGE HOURLY EARNINGS AND SKILL, OCTOBER, 1937

AVERAGE HOURLY EARNINGS (IN CENTS)	SKILLED	SEMI- SKILLED	UN- SKILLED	TOTAL
Under 37.5.....	.4	.7	2.2	.8
37.5 and under 42.5.....	.3	1.4	1.9	.9
42.5 and under 47.5.....	.5	1.0	3.9	1.2
47.5 and under 52.5.....	.5	2.5	9.3	2.7
52.5 and under 57.5.....	2.3	5.4	17.0	5.8
57.5 and under 62.5.....	4.1	10.3	20.6	8.8
62.5 and under 67.5.....	7.5	16.3	14.1	11.2
67.5 and under 72.5.....	7.5	14.2	15.6	10.8
72.5 and under 77.5.....	7.8	15.6	7.1	9.9
77.5 and under 82.5.....	10.7	11.5	4.4	9.8
82.5 and under 87.5.....	14.8	13.5	2.0	12.1
87.5 and under 92.5.....	10.7	3.6	1.0	7.0
92.5 and under 97.5.....	10.2	2.3	.5	6.2
97.5 and under 102.5.....	7.5	1.1	.3	4.4
102.5 and under 107.5.....	5.2	.2	..	2.9
107.5 and under 112.5.....	5.3	.4	.1	3.0
112.5 and under 125.0.....	3.5	..	..	1.8
125.0 and over	1.2	..	..	.7
Total.....	100.0	100.0	100.0	100.0

B

HOURLY EARNINGS RECEIVED BY EXPLOSIVES WORKERS IN THE UNITED STATES
ACCORDING TO GRADES OF SKILL, OCTOBER, 1937

GRADES OF SKILL	HOURLY EARNINGS RECEIVED IN CENTS		
	Middle Wage Received by Lower Half of Workers	Median	Middle Wage Received by Upper Half of Workers
Skilled	73.7	85.3	96.4
Semi-skilled	63.6	71.9	80.8
Unskilled	54.8	61.3	69.4

* Adapted from *Monthly Labor Review*, United States Department of Labor, Bureau of Labor Statistics, Vol. 47, No. 2 (August, 1938), pp. 383, 384.

Table 17 furnishes an example of a derived table in which some of the original data are presented along with the analysis. It appears rather complex but a study of the per cent columns reveals that only one form of analysis is being presented, namely the per cent of dwellings vacant in each ward for the various types of buildings.

ESTABLISHED PRACTICE IN THE CONSTRUCTION OF TABLES

The difference in purpose and content between primary and derived tables has been explained, and in general the rules for tabulation will apply to either kind. Derived or summary tables appear in print more frequently and they present the chief problems in table construction from the point of view of utility to the reader. The users of statistical analyses may be divided into two groups, those who will read tables and those who will not. As far as the second type of readers is concerned tabular matter might just as well be kept out of print. For their benefit it is necessary to point out in the text the most important information contained in any table. Those who will read tables would prefer to have the textual description omitted so that they can draw their own conclusions from the data presented. For the sake of this group the tables must be made as effective as possible.

Certain principles and practices which contribute to that end have become well established by usage and should generally be followed, although occasionally some deviation from customary procedure may increase the effectiveness of a table. In such cases it is more important to use good judgment than to follow rules slavishly.

Unity

The data contained in a table should pertain to one definite subject, should be confined to that subject, and the table should include whatever information is pertinent to a complete presentation of the subject.

In Primary Tables.—In a primary table there can be no question as to unity regardless of the degree of cross-classification if the subject is presented in terms of a single unit, as illustrated by Table 12, page 124, in chapter VII.

Other primary tables may contain information expressed in several units but with no sacrifice of unity due to the fact that significant ratios can be derived from combinations of the various sets of data. Table 18 is of this kind. The primary data given in the table are in three differ-

TABLE 17
A SAMPLE OF DWELLING PLACES WITH PER CENT VACANT, BY WARDS AND NUMBER OF FAMILIES ACCOMMODATED IN BUILDING,
BOTH RESIDENTIAL AND COMBINATION RESIDENTIAL AND BUSINESS BUILDINGS, BUFFALO, NEW YORK, 1936 *

Ward No.	DWELLING PLACES IN BUILDINGS ACCOMMODATING									
	One Family		Two Families		Three Families		Four Families		Over Four Families	
	No.	Per Cent Va- cant	No.	Per Cent Va- cant	No.	Per Cent Va- cant	No.	Per Cent Va- cant	No.	Per Cent Va- cant
1.....	394	2.3	382	0.0	45	0.0	4	0.0	...	...
2.....	476	0.2	348	0.0	15	0.0	..	..	10	0.0
3.....	273	1.8	438	0.0	9	0.0	4	0.0	...	...
4.....	190	19.5	320	2.2	150	1.3	72	0.0	45	0.0
5.....	206	8.7	368	1.4	222	0.5	152	0.0	100	0.0
6.....	197	14.7	410	0.0	138	0.0	40	0.0	...	...
7.....	728	1.2	542	0.7	90	0.0	40	0.0	15	0.0
8.....	230	13.9	316	1.9	156	2.6	96	3.1	78	3.8
9.....	230	2.6	492	1.6	222	2.3	104	0.0	26	7.7
10.....	316	3.8	424	0.0	72	0.0	8	0.0	6	0.0
11.....	240	5.0	450	0.7	225	0.4	52	0.0	12	0.0
12.....	309	4.9	306	0.3	63	1.6	24	4.2	21	0.0
13.....	302	4.3	540	0.0	63	1.6	32	9.4	32	0.0
14.....	314	1.5	542	0.0	69	0.0	28	0.0	5	0.0
15.....	470	1.9	582	0.3	35	3.0	8	0.0	...	...
16.....	323	2.2	372	0.0	45	3.0	8	0.0	12	0.0
17.....	687	1.5	406	0.0	27	3.7	4	0.0	...	...
18.....	535	5.0	280	1.4	9	0.0	16	0.0	8	12.5
19.....	319	6.9	544	1.5	18	0.0	12	0.0	24	0.0
20.....	327	4.6	312	4.8	21	0.0	8	0.0	31	0.0
21.....	719	9.0	808	0.1	150	0.7	72	1.4	11	0.0
22.....	370	8.6	432	2.3	105	0.0	28	0.0	36	0.0
23.....	173	13.9	288	5.2	30	0.0	8	0.0	78	5.1
24.....	218	7.8	262	4.6	48	0.0	52	0.0	171	8.8
25.....	272	5.5	266	2.3	81	0.0	60	0.0	340	5.3
26.....	242	10.7	420	3.8	78	0.0	40	0.0	55	25.5
27.....	155	31.6	250	0.4	171	1.8	140	2.1	311	32.5
Total.....	9,215	5.6	11,100	1.1	2,355	0.9	1,112	1.0	1,427	11.1
										25,209
										3.3

* From unpublished data compiled by the Bureau of Business and Social Research of the University of Buffalo.

ent units, number of telephones, number of messages, and number of dollars of operating income. However, all of them deal with the one subject, operations of the telephone company. Such ratios as number of messages per telephone, income per telephone, income per message of each type, as well as indexes showing relative changes in each unit or in each ratio, afford numerous possibilities for analysis.

In Derived Tables.—In a derived table simple units are replaced or supplemented by such measures as compound units, averages, and percentage relationships. These measures are usually based upon sev-

TABLE 18
OPERATING STATISTICS OF THE BELL TELEPHONE SYSTEM, 1931-36 *

YEAR	(1) NUMBER OF TELEPHONES	(2) (3) NUMBER OF MESSAGES		(4) (5) OPERATING INCOME		(6) TOTAL
		Local (000 omitted)	Toll and Long Distance (000 omitted)	Local Service	Toll Service	
1931	15,389,994	22,704,825	985,500	\$723,920,495	\$326,268,854	\$1,050,189,349
1932	13,793,229	21,525,558	823,866	670,736,747	263,147,955	933,884,702
1933	13,162,905	20,147,635	747,155	617,253,153	243,905,775	861,158,928
1934	13,378,103	20,676,520	781,830	607,676,275	258,691,363	866,367,638
1935	13,844,663	21,465,285	830,740	640,993,436	273,483,256	914,476,692
1936	14,453,552	22,869,510	911,340	665,152,512	306,238,511	971,391,023

* *Annual Reports of the American Telephone and Telegraph Company.*

eral separate sets of data but the derived table becomes a unified whole if all of the relationships included contribute to a single purpose.

An example of a compound unit is the "ton-mile," a measure of operating density in railroading. It represents one ton moved the distance of one mile and is derived from the two simple units "tons of freight" and "miles operated." Similarly in other lines of activity "man-hours," "dollar-years" and "foot-pounds" appear. Each of these is more complex than a simple unit yet each presents a single concept so that its use results in a unified table.

Table 17 is an example of a derived table in which percentages accompany data that have been classified in two directions. These percentages are based upon a third classification, number of dwellings occupied and vacant, the original data of which are not included. All of the information in the table contributes toward the one subject, percentage of vacancy in dwelling places. It has already been noted that a derived table loses its effectiveness if it presents too

many kinds of relationships. A number of additional sets of percentage relationships could be worked out from the original data of Table 17, such as percentage distribution of vacancy by wards or by type of building, but if any of these were included there would be a loss of unity and the resulting table would have no definite purpose.

Example of Lack of Unity.—Many tables appear in print which do not possess the unity found in Tables 17 and 18. An example of heterogeneous data presented in condensed form is shown in Table 19. The information on types of freight car loadings together with the Federal Reserve index forms a table complete in itself. "Pullman passengers carried" has nothing to do with freight car loadings nor with "financial statistics." "Canal traffic" on only two canals, and measured in two different kinds of tons, is possibly the most important information on the subject of water traffic, but more complete data should be given in a separate table, since there is no possible way of relating this information to that concerning freight car loadings. Tables that contain several sets of unrelated information or of related data expressed in units that are non-comparable are justifiable only in publications in which space-saving is a more important consideration than unity.

Complexity

The orders of classification employed in an investigation depend upon the nature of the data and the purpose for which they are collected. The extent to which it is necessary to study combinations of the several characteristics of the data will determine the degree of cross-classification required in tabulation.

Simple Classification.—The first order of classification, commonly referred to as a "one-way table," has been illustrated in Table 15. In each part of this table the prices of wheat are classified according to a single characteristic and no difficulty arises in their presentation. A derived table that contains a single classification follows the same form of construction.

Cross-Classification.—If classification is desired according to two characteristics simultaneously it will not serve the purpose merely to list the two separately. They must be cross-classified in a "two-way" table. This obviously requires listing in two directions, consequently one classification will appear horizontally and the other vertically, as in Figure 21. In this case the kinds of animals slaughtered are listed

TABLE 19
RAIL AND WATER TRAFFIC*

YEAR AND MONTH	FREIGHT CAR LOADINGS										FINANCIAL STATISTICS, CLASS I RAILWAYS			CANAL TRAFFIC			
	F. R. Index		Total	Coal and Coke	For- est Prod- ucts	Grain and Grain Prod- ucts	Live- stock	Mer- chan- dise l.c.l.	Ore	Mis- cel- lane- ous	FREIGHT CAR SURPLUS	PULL- MAN PAS- SEN- GERS CAR- RIED	Oper- ating Reve- nues	Net rail- way Op- erating Income	Net In- come	Sault Ste. Marie	Pan- ama
	Monthly aver- age, 1923-25 = 100										Thous- ands	Thousands of dollars	Thous. of short tons	Thous. of long tons			
	Unad- justed	Ad- justed															
1929: May....	107	107	1,032.6	166.9	68.6	37.8	25.0	255.1	72.7	406.5	223	2,590	531,823	101,332		13,930	1,206
1932: May....	51	52	500.4	73.4	17.9	27.9	16.1	174.5	2.6	188.0	751	1,270	251,922	11,666	29,791	1,568	662
1933: May....	54	55	532.3	85.6	21.9	35.8	15.8	161.5	8.7	202.9	553	951	255,241	41,043	2,969	3,490	783
1934: May....	63	64	605.2	111.9	25.1	28.1	16.0	159.3	25.8	239.0	355	1,122	282,039	39,699	3,337	5,745	1,008
1935: May....	60	61	579.8	112.4	24.3	25.0	12.2	153.3	28.7	223.8	305	1,146	279,528	39,599	4,668	5,985	938
1936: May....	71	72	670.2	117.6	32.4	30.7	11.6	158.0	42.3	277.5	185	1,295	320,926	41,797	2,308	8,710	940

FORM FOR FIVE-WAY CROSS-CLASSIFICATION

NUMBER OF CATTLE, CALVES, HOGS AND SHEEP SLAUGHTERED UNDER FEDERAL AND CITY INSPECTION BY THREE MEAT PACKERS IN 1936 AND 1937, SUBDIVIDED BY GRADE OF MEAT

[illegible]

resulting in a "four-way" table. Figure 23 in which grade of meat has been added, and Figure 24 adding a time classification, illustrate respectively the method of combining four and five orders of classification in a single table. Additional classifications could be introduced if there were others pertinent to the data. Inspection of Figures 23 and 24 shows, however, that they contain too much information to be comprehended easily. Whenever the further subdivision of data leads to tables which are too complex to be read easily, it is preferable to increase the number of tables. Do not spend time devising ways of presenting multiple classifications in a single table; make two or more tables instead. A fairly good, though not universal, rule is to confine a table to three classifications if it is being made for publication.

From the point of view of construction the addition of a set of percentage relationships to the original data increases the complexity but does not add another order of classification. Thus if the percentage of animals slaughtered by each packer were required in Figure 21, it would be necessary to introduce the subheadings "Number" and "Per Cent Distribution," under each type of animal in the caption. It would then have the appearance of a three-way table although only a two-way classification.

Clarity

The reader's ability to grasp the content and significance of a table depends primarily upon the clarity of wording in every part of the table. Careful attention must be given to the phraseology of the title and all headings, and to the inclusion of any necessary notes of explanation and reference.

Title.—The first essential is a title which in the simplest form will tell what is in the table. If several lines are required to describe the contents, a brief title explaining the major characteristics can be used with a subtitle in smaller print giving the more detailed description.

The title should clearly name the unit or units in which the data are being presented including all the limitations on the data. These limitations usually include time, space, and exact specifications of the units employed. If a presentation of some derived relationship is the main purpose of the table that relationship should be stated with equal precision. The methods of classification used should also be specifically indicated especially in studies of limited scope when it is only this latter qualification that distinguishes one table from another. In such

cases it is not necessary to repeat in each table the general description common to all of them.

Stated briefly, if the title definitely answers the questions, what, where, when, and how, it is probably adequate. No better guide can be found for the correct and specific wording of titles than the usage in the United States Bureau of Census publications. Such titles as "Prime Movers, Motors and Generators, by Number and Rated Capacity, for Establishments Classified According to Number of Wage Earners Employed: 1929" or "Percentage of Homes Owned and Rented, by Color and Nativity of Head of Family, for the United States: 1930, 1920, 1900 and 1890" illustrate how the various phrases needed in a complicated title may be best arranged to give a clear idea of the content of the table.

Headings.—Every part of the table requires a heading. This includes general headings for caption and stub, for each order of classification, and for each of the separate classes. Clarity and brevity are the chief considerations in wording these headings. They must be complete enough to serve as accurate guides to the data, although lack of space usually requires that they be worded as briefly as possible. There should be no unnecessary repetition of information that has already been given in the title. Each main heading should include whatever detail applies to all of its subheadings so that the latter which are the most crowded of all may be very short. The mechanics of arranging these headings contributes to the effectiveness of the table and will be discussed later in the chapter.

In referring to the data in a table it is convenient to be able to designate the columns by number. In Figure 24 the numbers are placed at the extreme top of the column headings. As an alternative the numbers can be placed just above the line separating the headings from the data as shown in Figure 23. Sometimes the horizontal rows of a table are also numbered but this practice is less common. The numbering of columns or rows is not a requirement but merely a convenience to be used whenever it will facilitate the description of the table in the text or reference to it in subsequent tables. The stub itself is ordinarily not numbered as a column.

/ Another common practice that aids in reading a table is the insertion of the unit directly over the columns to which it refers, as illustrated in Table 19. Thus the data of certain columns are clearly designated as thousands of cars, thousands (of passengers), thousands of

dollars, and thousands of tons, respectively. In the index number column "Monthly average, 1923-25 = 100" is inserted in a similar position. When the headings are in the stub the units are stated along with each item or in a separate column, as in Table 51, page 289.

/ *Footnotes.*—Anything in a table which cannot be understood by the reader from the title and headings should be explained in one or more footnotes. These footnotes should contain statements concerning figures that are missing, preliminary or revised, and explanations concerning any unusual figures or other features of the table that are not self-explanatory. A study of tables appearing in print will provide multiple illustrations of the use of footnotes. Table 27, page 222, is an excellent illustration.

/ *References.*—A table should always give exact reference to the source or sources from which the data were taken. Three advantages grow out of such citations: (1) The reader is given a sound basis for evaluating the data. (2) Readers who wish to obtain other data similar to those appearing in the table are able to do so. (3) The author of the table insures himself against the inconsistency of source books. For example, the data for wheat production will depend entirely upon which issue of *Agricultural Statistics* one happens to use. If a table contains data that have not been published previously, this fact should be stated in a note including the name of the collecting agency. In general, the use of exact references is a method of guarding against the charge of inaccuracy.

Arrangement

/ The arrangement of a table on the page, the arrangement of data in the table and the choice of ruling, spacing, and type face contribute to the effectiveness.

Fitting the Table to the Page.—The limitations of the size of the printed page determine the form of tables to a large extent, hence the real problem of arrangement is to fit the table to the page so that it will be effective in that setting. One of the most important features is symmetry with respect to the margins and binding of the page. The table should be planned to read from left to right. Tables which must be read from the side of the page, tables which cover two facing pages, and tables which must be unfolded either sideways or vertically are occasionally necessary, but they should be used only when no combination of smaller tables will serve as well.

The proportions of the page may also determine which headings to use as stub and which as caption. In order that the height of a table may exceed the width, a one-way table is usually arranged vertically and in a cross-classification the longer list of items will ordinarily appear in the stub. Length of wording of the headings is another factor to consider. It is better to use the longer wording in the stub if possible since too many words crowded into a narrow column heading are very hard to read.

When additional classifications are introduced into either the caption or the stub, the new headings may be arranged in subordinate positions as in Figures 22, 23 and 24, or the table may be rearranged to make the former classification subordinate to the new one. In all such cases the question will arise of arranging columns and rows so as to emphasize significant relationships. The chief consideration in Figure 24 is comparison between 1936 and 1937, consequently the time classification has been arranged in pairs of adjacent columns. If this were not the case, time could be made a main classification leaving the types of inspection in adjacent columns, as follows:

CATTLE					
1936			1937		
Federal	City	Total	Federal	City	Total

Order of Items.—The three types of classification according to time, space, and attributes have been discussed. A classification in any of these categories frequently results in a large number of subdivisions or classes, which necessitates the introduction of a definite order of arrangement.

Time classifications follow the natural order of the events represented. It is only when the major emphasis falls on the most recent events that a reverse time order is used.

Spatial classifications sometimes follow the order of geographical proximity as when the main subdivisions of the United States are given from northeast to southwest. The New England States are frequently named in the order familiar to everyone, Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, and Connecticut, but ordinarily a list of any length is most usable if arranged in alphabetical order. Size or importance is also occasionally the basis for spatial

classification. The method used must be determined according to the nature of the data and the reader's familiarity with the criterion selected.

Attributes of quality lead to great variety of tabular arrangement. These may be listed according to importance or in some other order familiar to the reader, but again for comparability and ease in locating any given item the alphabetical arrangement is preferable.¹

An example of time arrangement in natural order appears in the stub of Table 19, page 156, and in the same table the classification of types of freight car loadings is alphabetical. It should be noted that an item such as "miscellaneous" or "other" is placed at the end of the list in any order of arrangement. Table 20 shows three examples of other orders that may be followed.

In Table 20-A the data in a spatial classification are listed according to the size of the city in which they occur. Tables 20-B and 20-C show two possible arrangements of one set of data in an attribute classification, one being in order of size and the other an arbitrary arrangement. In the former it is the size of the per cents, i.e., the data that are being tabulated, that is used as a basis for the arrangement of the classes, whereas in 20-A the determining size was that of the geographical classes themselves rather than the amount of tax receipts. The order of 20-C is a combination of form of organization, importance, and respectability. That is, commercial banks, industrial banks, and personal finance companies are privately owned corporations operated for profit. The next four are co-operative plans usually conducted on a small scale. The pawnbrokers and unlicensed lenders are the most important members of the group but are not on the same plane of respectability as the others because of the questionable business methods sometimes employed.

Ruling, Spacing, and Type Face.—These are devices for increasing the effectiveness of a table by concentrating emphasis on important entries and by relieving the monotonous appearance of figures in rows and columns. Whenever rulings aid the reader in understanding the classifications and subclassifications of a table they should be used. Double and triple rulings are not necessary since equal effectiveness can be achieved by using a single heavier line to separate major divisions in a table. It will be observed that in Figures 23 and 24, pages 158 and 159, every column is separated from the next by rulings, but that only the main groups of rows are so separated. In many printed

TABLE 20

EXAMPLES OF TABLE ARRANGEMENTS

A			B	
SPATIAL DISTRIBUTION ACCORDING TO SIZE OF CHARACTERISTIC			ATTRIBUTE DISTRIBUTION ACCORDING TO SIZE OF DATA	
TOTAL TAX RECEIPTS OF LARGE CITIES, 1935*			PERCENTAGE DISTRIBUTION OF SMALL LOAN BUSINESS DONE BY VARIOUS LENDING AGENCIES†	
CITY NUMBER IN ORDER OF SIZE	CITY	TAX RECEIPTS (000,000 omitted)	TYPE OF LENDING AGENCY	PERCENTAGE OF SMALL LOAN BUSINESS
1	New York	\$586	Unlicensed lenders.....	28.9
2	Chicago	209	Pawnbrokers	23.2
3	Philadelphia	90	Personal finance companies	19.3
4	Detroit	82	Industrial banks	13.9
5	Los Angeles	57	Commercial banks.....	7.3
6	Cleveland	42	Credit unions.....	2.4
7	St. Louis	31	Remedial loan societies....	2.3
8	Baltimore	34	Axias	1.9
9	Boston	65	Employers' plans.....	.8
10	Pittsburgh	42	Total	100.0
11	San Francisco	31	C	
12	Milwaukee	34	ATTRIBUTE DISTRIBUTION ACCORDING TO ARBITRARY ARRANGEMENT	
13	Buffalo	33	PERCENTAGE DISTRIBUTION OF SMALL LOAN BUSINESS DONE BY VARIOUS LENDING AGENCIES†	
14	Washington	29	TYPE OF LENDING AGENCY	PERCENTAGE OF SMALL LOAN BUSINESS
15	Minneapolis	23	Commercial banks.....	7.3
16	New Orleans	19	Industrial banks.....	13.9
17	Cincinnati	19	Personal finance companies..	19.3
18	Newark	32	Employers' plans.....	.8
19	Kansas City	18	Credit unions.....	2.4
20	Seattle etc.	15	Remedial loan societies....	2.3
			Axias	1.9
			Pawnbrokers	23.2
			Unlicensed lenders.....	28.9
			Total	100.0

* *Statistical Abstract*, U. S. Department of Commerce, 1937, p. 218.† Evans Clark, *Financing the Consumer* (New York: Harper & Bros., 1933), p. 30.

tables there are no horizontal rulings but the separation between the rows and groups is accomplished by appropriate spacing and by successive indentations of items to indicate various degrees of subclassification. Bold face type, larger type, and italics are frequently used to set off totals or percentages from the other data or to emphasize important items. The use of these devices is well illustrated in the tables

presented in the first section of each monthly issue of the *Survey of Current Business*.

Totals

What Totals To Include.—A table is not complete unless it includes whatever totals and subtotals are required to summarize the data presented, but this does not mean that every row and column must be totaled. A total implies that the same unit is used in all of the classes added and that the several classes taken together form a homogeneous whole.

This principle can be explained by reference to Figure 24. The four kinds of animals slaughtered, cattle, calves, hogs, and sheep, have not been added together because a total would imply, for example, that a beef carcass means the same thing as a hog carcass in terms of meat and meat products. The time classification has not been totaled, since a total made up of two years' observations would be purely arbitrary. The emphasis in a time comparison such as this is usually on the relation of the several different periods to each other. However, in case the entire period covered in a table represents a genuine total, such as a year's production resulting from the sum of the production for 12 months, a total for the year would have significance.

On the other hand a total for each kind of animal by inspecting agency is included because the classification represents the separation of a whole into comparable parts and because the total is a production figure of value for comparisons in the other classifications. Likewise three sets of totals have been computed vertically: (1) the subtotal for each company, (2) the subtotal of each grade of meat, and (3) the total of both grades of meat for all three companies. It will be noted that the selection of subclassification in the stub of Figure 24 has resulted in bringing together in the last section the subtotals of each grade of meat, whereas the subtotals for each company are separated from each other. Comparisons of the latter can easily be made since the *company subtotals* and the *total of all companies* are printed in italics.

The preceding examples dealt entirely with classifications in which the units were counted objects and included totals wherever pertinent. In classifications of counted objects which are not parts of a total, sometimes an adjusted total may be used, but in general no total should be included. Distinct from these are classifications of rates or prices

as Table 15, page 149, in which averages should take the place of totals whenever a summary figure is required to make the table complete. A derived table may include totals, averages, or neither, according to the nature of the data and the purpose for which it is constructed.

Position of Totals.—The natural sequence of reading is properly accommodated by placing totals at the foot of the columns and at the right of the rows. Statistical practice may reverse this, placing totals at the top and left when they are more important than the individual items. Both positions are in common use. This question must be decided for each table in terms of what the maker believes will best serve his purpose. Briefly, the usage is: totals at the top and left or at the bottom and right; but not at the top and right, nor at the bottom and left. Subtotals follow the same usage in any given table, that is, if the totals are at the top and left of the table the subtotals will also appear at the top and left of the items from which they are computed.

Significant Figures

There are two aspects of the subject of significant figures in a table. The first relates to the number of significant figures which need to be retained for accuracy, while the second relates to a total and its parts.

Number of Figures Retained.—Tables are often unnecessarily encumbered by retaining all of the digits in numbers running to millions or even billions. A good example of this is Table 18, page 154, in which too many digits have been retained purposely for discussion at this point. It was stated in chapter II² that four significant figures are adequate for statistical work. The unit for telephones could therefore be changed to 10,000 telephones, but customarily units are used in thousands, millions, or billions. In accordance with this practice the number of telephones should be expressed in thousands, making the figures accurate to five digits. The column showing the number of toll calls should be expressed in units of one million calls accurate to one decimal place, giving four significant figures. The entire table has been reproduced in corrected form as Table 21. The revised form is more effective and sufficiently accurate for statistical purposes.

The general rule is, retain only four or at most six significant figures in the tabular presentation of data, but in all cases indicate the size of the unit used either by showing the number of digits omitted from

² p. 27.

TABLE 21
REVISED FORM OF TABLE 18

YEAR	(1)	(2)	(3)	(4)	(5)	(6)
	NUMBER OF TELEPHONES (000 omitted)	NUMBER OF MESSAGES (000,000 omitted)		OPERATING INCOME (000,000 omitted)		
		LOCAL	TOLL AND LONG DISTANCE	LOCAL SERVICE	TOLL SERVICE	TOTAL
1931.....	15,390	22,705	985.5	\$723.9	\$326.3	\$1,050.2
1932.....	13,793	21,526	823.9	670.7	263.1	933.9
1933.....	13,163	20,148	747.2	617.3	243.9	861.2
1934.....	13,378	20,677	781.8	607.7	258.7	866.4
1935.....	13,845	21,465	830.7	641.0	273.5	914.5
1936.....	14,454	22,870	911.3	665.2	306.2	971.4

each number or by using the expressions "in thousands" or "in millions," etc. Note that in column 3 "(000,000 omitted)" refers to the number of digits dropped between the original decimal place and the newly established one, regardless of the fact that the data as written are accurate to hundred thousands,

Rounding Off Totals.—A different question arises when the table consists of a total and its parts. The entire table should be made up from the original data and each item then rounded off separately, as the data in Table 21 were rounded off from Table 18. As a result the sum of the individual items as they appear in the rounded-off table may be either greater or less than the rounded-off total of the original data. One such instance may be seen in Table 21. For 1932 the total operating income, column 6, does not correspond exactly to the sum of its parts in columns 4 and 5, although reference to Table 18 will show that no error has been made either in the addition or in rounding off any of the three figures.

The percentage distribution is a particular case of the part to total relation which requires further explanation. For example, in Table 22 the percentage distribution of the interest-bearing debt should add to 100 per cent whether it is carried to one decimal place or two, because the total debt is being distributed and the sum of the parts must be equal to the whole. However, the exact sum of column 2 is 99.99 and of column 3 is 100.1. The discrepancy arises from rounding off the last decimal place in the computations. No theoretical question is involved but merely the practical one of the best method of expressing the total in the table. Since the total is exactly 100 per cent, in any such case of apparent discrepancy it should be written to one less sig-

TABLE 22

THE INTEREST-BEARING DEBT OF THE UNITED STATES TREASURY ON APRIL 30, 1937,
ACCORDING TO TYPE OF OBLIGATION, AMOUNTS AND PERCENTAGE DISTRIBUTION*

TYPE OF OBLIGATION	(1) AMOUNT OUTSTANDING (000,000 omitted)	(2) PERCENTAGE DISTRIBUTION (two decimals)	(3) PERCENTAGE DISTRIBUTION (one decimal)
General bonds.....	20,133.7	58.70	58.7
U. S. savings bonds	755.5	2.20	2.2
Adjusted service bonds.....	409.6	1.19	1.2
Treasury notes	10,377.4	30.26	30.3
Certificates of indebtedness.....	268.7	.78	.8
Treasury bills.....	2,353.2	6.86	6.9
Total	34,298.1	100.0	100.

* *Statement of the Public Debt of the United States, April 30, 1937, Treasury Department.*

nificant figure than appears in the individual per cents. Thus the total of the first percentage distribution has been written 100.0 and the second 100.

Checking

Any possible errors in a table should be guarded against by complete checking. There may be errors in the numerical content due to mistakes in addition or other computations, or the validity of the table as a whole may be impaired by errors in judgment with regard to the general plan, items included, or details of wording and arrangement.

Accuracy of Numerical Content.—Checking of computations should be a routine step in table construction. Whenever any part is totaled the addition must be checked, preferably on a separate outline form. If there are horizontal totals and vertical totals as in Table 17, page 153, the grand total (25,209) that is common to the two sets of totals must be checked in both directions. If the form is similar to Figure 24, page 159, all parts of each subtotal must be checked as well as the grand total. It cannot safely be assumed that if the totals of a table check in one direction or if both sets of subtotals check with the grand total there are no mistakes in the figures. Similar precautions must be taken in the case of every computation in every step of table preparation, especially apparently simple operations that are done mentally such as multiplying by 2 or dividing by 25. Each step should be checked before the next one is taken and corrections of errors should be rechecked.

Validity.—Errors of judgment may occur in the planning of tables or during their construction. Checking by an experienced person who

has an adequate knowledge of the background of the subject as well as experience in table construction is required if such errors are to be avoided. The types of errors that could have been corrected by careful checking are illustrated in Table 23.

TABLE 23
LIVE-STOCK EXPORTATION, 1929-32
(1,000 head)

YEAR	1928		1950		1931		1932	
		PER CENT		PER CENT		PER CENT		PER CENT
Hogs, live	1,279	88	654	52	355	22	179	13
Hogs, butchered	47	3	117	9	191	11	16	1
Bacon	116	8	410	32	920	56	1,008	74
Ham and other products, pieces..	8	1	76	7	179	11	160	12
Total	1,450	100	1,257	100	1,645	100	1,363	100

The title fails to name the country exporting. It indicates all kinds of livestock but the only live animal included in the table is hogs. Butchered hogs can scarcely be called livestock and certainly bacon and hams are not. The title names the years 1929-32 which might mean either annually or a total for the period. However, the headings do not agree with either, even disregarding the obvious typographical error in printing 1950. Since there are at least three different units used in the table—head of hogs, pieces of ham, and bacon in some unnamed unit—the unit “1,000 head” should not have been stated along with the title. It should have read “All figures in thousands.” The caption heading “Year” is misplaced and the stub has no heading. The caption subheadings should read “Number” and “Percentage Distribution.” However, the most fundamental error is in the presentation of totals and percentage distributions for items that are expressed in non-comparable units and that in no sense make up a total having any meaning.

✓ TABULAR FORMS

In the conduct of routine statistical work the tabulation of data becomes a continuous process of recording the same types of data daily, weekly, monthly, or annually as the case may be. For this purpose the preparation of standard forms in quantity saves time and promotes uniformity of records. These forms must be carefully planned and drawn up, hence they are excellent illustrations of the

use of the principles of tabulation. Frequently special adaptations must be made to facilitate the recording of particular types of data, a circumstance which increases the desirability of studying such forms. On succeeding pages selected forms used by government agencies and by a business concern are presented with brief discussion.

Government Forms

Three forms used in the price section of the crop reporting service of the Department of Agriculture are presented as examples of the blanks prepared for recording external data.

Figure 25 is used principally for summarizing the data on prices received by farmers by months and by commodities. The printed headings indicate the following: Column 1, weights for computing weighted average prices for the United States; column 2, straight or unweighted average prices reported by correspondents in each State; column 3, average prices reported by corre-

FIGURE 25

DEPARTMENT OF AGRICULTURE FORM C. E. 1-128

[illegible]

sheet. These forms are also bound together in books in which all States, geographic divisions and the United States are represented.³

These forms are so planned that they can be used for several different kinds of tabulations. For example, Figure 27 is also used to record the individual crop reports from which the state averages of column 2 of Figure 25 are computed.

Business Forms

The routine tabulation by business concerns of data concerning their own operations differs from the record keeping of government agencies mainly in the type of data tabulated. This difference leads to the use of forms which are quite distinct from those previously presented, and which are specialized to meet the needs of the particular concern using them. It follows that as many forms could be presented as there are business concerns, but those used by a single concern will illustrate some important uses and adaptations.

The forms shown on succeeding pages are used in the routine statistical work of the Eastman Kodak Company of Rochester, New York.⁴ This company divides the year into 13 four-week periods, hence all of the forms which follow are filled out 13 times each year.

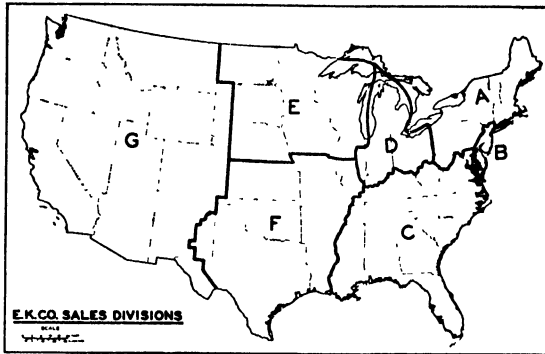
Figure 28.—The "Comparison of Sales by Divisions" is a form prepared primarily for the use of executives. It is the usual type of two-way table intended to be read in either direction, i.e., the percentage of increase or decrease of sales of any of the products can be compared from one division to another or the percentage of increase or decrease of sales of different products in any division can be compared. The percentage of increase or decrease of sales by divisions in next to the last column is not obtained from the preceding columns but by comparing sales in dollars with the corresponding figure for the preceding year. These percentages can be compared with the percentage of change in bank debits given in the last column.⁵ The preparation of the percentage of change of bank debits requires considerable work because the divisions of the country used by Eastman Kodak

³ The explanation of the use of these forms was supplied by Mr. Roger F. Hale, Agricultural Statistician, Division of Crop and Livestock Estimates, Bureau of Agricultural Economics, United States Department of Agriculture, Washington, D. C.

⁴ They are presented with permission of the Eastman Kodak Company and are made available through the courtesy of Mr. A. H. Robinson, Assistant-Treasurer.

⁵ The use of bank debits as an indicator of business activity is discussed in chapter XX.

FIGURE 28
COMPARISON OF SALES BY DIVISIONS



PERCENTAGE CHANGE FROM PREVIOUS YEAR

Black figures are increases, red figures are decreases

DIVISION	AMATEUR CAMERAS	AMATEUR FILM	C K PRODUCTS	PROF FILM & PLATES	PAPERS	CHEMICALS	X RAY & DENT FILM	TOTAL*	BANK DEBITS
A									
B									
C									
D									
E									
F									
G									
TOTAL U S									

* Excludes sales not common to all divisions

Company as shown on the map do not coincide with Federal Reserve Districts for which the percentages of change in bank debits are available in print.

Features of this form are the ruling, the use of black and red ink to distinguish increases and decreases and the exclusion from the total column of sales not common to all divisions. For example, if there were some divisions in which "chemicals" were not sold, the sales of this product would be excluded from all divisions in computing the percentage of change of total sales.

Figure 29.—The line above the double ruling might read "Report for Fifth Period Ending May 20, 1939." The record of

LOST TIME REPORT

EASTMAN KODAK COMPANY

LOST TIME REPORT										SUMMARY OF TOTAL EMPLOYEES				
EASTMAN KODAK COMPANY										Avg. for Period		At End of Period		
Report for										Male (Excl. Supts., etc.)		Female (Excl. Supts., etc.)		
Period Ending.....										Dept. Supt. & Asst.		Supt., Main Office Dept.		
K O 465 K P 23726										Holds		Total		
LOST TIME REPORT—GENERAL EMPLOYEES														
Number of Employees and Times Absent														
		Male		Female		Total		Time Lost—Hour		Total				
Code No.	REASON	Persons	Times Absent	Persons	Times Absent	Persons	Times Absent	Male	Female					
1	With Permission													
2	Without Permission													
3	Sick Work													
4	Sickness													
	8 hours and less													
	Over 8 hours to 40 hours													
	Over 40 hours to 26 weeks													
5	Accident During Work													
6	Accident After Work													
7	Vacation Paid For													
8	Discipline													
9	Excess Over 40 Hours													
Total Lost Time (Excl. Vacations and Holidays)										Vacation Pay		Average Hourly Rate of Earnings		
Total Lost Time (Excl. Holidays)										Incl. in Total Payroll		Total Payroll		
Total General Employees (Average for Period)										Payroll Time		Total		
Total Actual Working		Actual Overtime Hrs. Worked (Incl. in Total Hours Worked) Excludes sick benefit double time was paid		Total Actual Payroll Incl. in Total Payroll Double Rate		Overtime Pay Inc. in Total Payroll Double Rate		Vacation Pay Incl. in Total Payroll		Average Hourly Rate of Earnings				
Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female			
Total		Total		Total		Total		Total		Total				
*Excludes sick benefit payments and compensation payments														

LOST TIME REPORT—NO LOST OR OVERTIME EMPLOYES (All figures in this section exclude Manager, General Supt., Supt., etc.)									
Code No.	REASON	Number of Employees and Times Absent					Time Lost—Hours		
		Male		Female		Total	Male	Female	Total
		Persons	Times Absent	Persons	Times Absent				
1	With Permission								
2	Without Permission								
3	Sick Work								
4	Sickness								
	8 hours and less								
	Over 8 hours to 40 hours								
	Over 40 hours to 26 weeks								
5	Accident During Work								
6	Accident After Work								
7	Vacation Paid For								
8	Discipline								
9	Excess Over 40 Hours								
10	Holiday								
Grand Total Lost Time									
Total Lost Time Not Paid For									
(Incl. in above Grand Total)									
Total Lost Time Paid For									
(Incl. in above Grand Total)									
Total No Lost or Overtime Employees (Excl. Supts., etc.) (Average for Period)									
Actual Total Regular Hours Worked (Excl. Supts., etc.)		Male		Female		Total		Actual Total Payroll** Incl. Pay for Vacations, Holidays, Sick Benefits, etc. (Incl. Supts., etc.)	
		Male		Female		Total		Male	
		Female		Total				Female	
		Total						Total	
**Excludes compensation payments.									
For Explanations See Reverse Side.									

(All figures in this section exclude Manager, General Supt , Supt., etc.)

Number of Employees and Times Absent

[illegible]

1	With Permission	8 hours and less Over 8 hours to 40 hours Over 40 hours to 26 weeks Accident During Work Accident After Work Vacation Paid For Discipline Excess Over 40 Hours Holiday
2	Without Permission	
3	Slack Work	
4	Sickness	
5		
6		
7		
8		
9		
10		

Grand Total Lost Time
Total Lost Time Not Paid For
(Incl. in above Grand Total)
Total Lost Time Paid For
(Incl. in above Grand Total)

**Total No Lost or Overtime Employees (Excl. Supts., etc.)
(Average for Period)**

Actual Total Regular Hours (Excl Supra, etc.)	Actual Total Payroll** Total Pay for Vacation, Sick Leave, Sick Benefits, etc. (Excl Supra, etc.)	Average Hourly Rate of Earnings
Male	Male	Male
Female	Female	Female
Total	Total	Total

**Excludes compensation payments.

For Explanations See Reverse Side.

FIGURE 30

INSTRUCTIONS ON THE REVERSE SIDE OF LOST TIME REPORT

1. Reportable Employees
 - (a) In "Summary of Total Employees" the figures on the total for Manager, General Superintendents, Departmental Superintendents and Assistant Superintendents, and Main Office Department Heads should be indicated in a separate item.
 - (b) The "Average for Period" figures are averages of the four figures for the employment at the end of each week of the period.
 - (c) General—Under the section on "General Employees" are to be reported all employees not on a "No Lost or Overtime" basis. The total "General Employees" figures should be the average for the period.
 - (d) No Lost or Overtime—Under the section on "No Lost or Overtime Employees" are to be reported all employees except "General Employees." Exclude the Manager, General Superintendents, Departmental Superintendents and Assistant Superintendents, and Main Office Department Heads. The total "No Lost or Overtime Employees" figures should be the average for the period.
2. All absences of one hour or more are to be reported. Absences for any cause except accidents are not to be reported for a period exceeding 26 weeks for any one employee.
3. Absences with Permission (Code No. 1)
 - (a) All absences with permission other than for slack work, illness, accidents, vacations paid for, holidays, discipline, or excess over 40 hours should be indicated in one item for both "General Employees" and "No Lost or Overtime Employees."
4. Absences without Permission (Code No. 2)
 - (a) Include all time lost without permission and when no explanation is given.
5. Slack Work absence (Code No. 3) should be indicated for both "General" and "No Lost or Overtime Employees."
6. All reportable absences on account of illness (Code No. 4) are to be reported in the following manner (this applies to both "General" and "No Lost or Overtime Employees").
 - (a) 8 hours or less.
 - (b) More than 8 hours, but not exceeding 40 hours.
 - (c) More than 40 hours, but not exceeding 26 weeks.
7. Absences on account of accident and injuries (Codes No. 5 and No. 6) should be reported until:
 - (a) Employee has returned to work.
 - (b) Employee has left employ of Company.
 - (c) It has been decided to be a case of permanent disability.
8. Vacations Paid For (Code No. 7)
 - (a) Under this heading include all time granted for annual vacations which is paid for. These figures should be given for both "General" and "No Lost or Overtime Employees."
9. Absence for discipline (Code No. 8) should be indicated for both "General" and "No Lost or Overtime Employees."
10. Absence for "Excess Time Over 40 Hours" (Code No. 9) should be indicated for both "General" and "No Lost or Overtime Employees."

FIGURE 30—(Continued)

11. **Holidays (Code No. 10)**
Absences for holidays should not be reported for "General Employees," inasmuch as they are generally not paid for but at the same time generally granted. Absence for holidays for "No Lost or Overtime Employees" should be indicated inasmuch as they are generally paid for.
12. Under the "General Employees" section, the total Lost Time figures should be split to show the total including "Vacation Paid For" and excluding "Vacation Paid For." In both cases the totals exclude holiday absences.
13. Under the "No Lost or Overtime Employees" section, the total Lost Time figures are split three ways. The Grand Total of Lost Time for all reasons is shown. The total Lost Time Not Paid For is indicated separately, and the total Lost Time Paid For is another separation. This breakdown is given on the totals because it is not readily available on the individual items of various reasons for absences.
14. For the "General Employees" division, the Hours Worked figures should be the actual hours on the job, including the overtime hours worked but excluding lost time for all reasons.
 - (a) The overtime paid for on a time and one-half basis should be shown separately in terms of just the actual hours worked. This figure is a portion of the total Hours Worked
 - (b) The overtime paid for on a double time basis should be shown separately in terms of just the actual hours worked. This figure is a portion of the total Hours Worked.
15. For the "No Lost or Overtime Employees" division, the total Regular Hours Worked should be the total hours paid for less all lost time hours paid for. These figures should exclude Superintendents, etc.
16. For the "General Employees" division, the Total Actual Payroll should be the total pay, including vacation pay and overtime pay but will exclude any sick benefit payments or compensation payments.
 - (a) The overtime pay for hours paid for on a one and a half times basis should be indicated separately. This figure is a portion of the Total Payroll.
 - (b) The overtime pay for hours paid for on a double time basis should be indicated separately. This figure is a portion of the Total Payroll.
 - (c) The Vacation Pay should be indicated separately. This figure is also a portion of the Total Payroll indicated.
17. For the "No Lost or Overtime Employees," the Total Actual Payroll should be the total pay, including vacation pay, holiday pay, sick benefit payments, and payments for all "lost time for any reason paid for," except compensation payments. These figures should exclude Superintendents, etc.
18. The *Average Hourly Rate of Earnings* should be arrived at as follows:
 - (a) For "General Employees," the "Total Actual Payroll, excluding Vacation Pay" should be divided by "The total actual hours worked including overtime hours plus the bonus hours paid for on the overtime" (that is, take time and a half basis hours and add 50% and take double time hours and add 100%).
 - (b) For "No Lost or Overtime Employees" the "total actual payroll, including vacation pay, holiday pay, sick benefit pay, etc." should be divided by "Total Regular Hours Worked plus all lost time hours paid for."
19. For all items of information, the figures for Male and Female Employees should be shown separately.

total employment is placed at the top of the sheet but in the remainder of the form the staff is divided into "General Employees" and "No Lost or Overtime Employees." The reasons for lost time at the left give a detailed view of what caused employees to be absent from work and the length of absence. A summary of actual hours worked, hourly rate of earnings, and payroll is included for each of the two types of employment.

The features of this form are the spacing, variation in type, judicious use of ruling, and the inclusion of a transfer code.

Figure 30.—A person unfamiliar with the organization and operations of the Eastman Kodak Company would experience considerable difficulty in filling out the form of Figure 29 because of the technical use of terms and the need for explanation of the method of computing averages and other summary figures. But variations in usage would also occur among those familiar with the form, if uniform interpretations of terms and computations were not provided, hence the instructions for filling out the form are printed on the back. These have been reproduced as Figure 30.

The explanations are written for the guidance of persons thoroughly familiar with the way the company is organized and operated; therefore much is omitted that would have to be stated in the instructions accompanying a similar form used in external statistical work. For example, the definition of reportable employees appears to be ambiguous in stating that "General Employees" excludes "No Lost or Overtime Employees" while "No Lost or Overtime Employees" excludes "General Employees." However, the persons who use this form understand that "General Employees" are those who work on a piece rate or hourly rate basis, and "No Lost or Overtime Employees" are those who work on a fixed weekly wage basis. Hence the instructions concerning reportable employees mean that regardless of the type of work performed by a particular individual during a given payroll period he is to be reported according to his permanent status either as a time worker or a salaried worker.

Figure 31.—The Labor Turnover Report like the Lost-Time Report is filled out for each works and each four-week period. The form is largely self-explanatory although it calls for considerable detail. The report requires information on three subjects: (1) the number employed, (2) the number entering, and (3) the number leaving, but the emphasis is placed on an analysis of the exits. The primary sum-

FIGURE 31
LABOR TURNOVER REPORT

SUMMARY				REASONS FOR LEAVING			
193							

mary figure is the "Net Turnover Per Cent" obtained by dividing "Net Number Leaving" by the "Average Number of Employees for the Period."

The percentage distribution of "Total Exits" according to length of service provides information concerning the "employee plant age" at which employment is most likely to be terminated.⁶

The analysis of reasons for leaving, on the right-hand side of the form, is intended to provide detailed information concerning the underlying causes of labor turnover. Continuous study of these reports aids management in two ways: (1) the personnel manager secures a background of knowledge of what types of persons are most likely to become permanent employees and (2) over-all management is able to detect unsatisfactory conditions that are causing a large separation ratio in any part of the organization.

Conclusion

The forms appearing in this section are not intended to be representative of all the prepared tables used in routine statistical work. The purpose is merely to present a few examples to show how the principles of tabulation are employed in practical work. The outstanding feature of all of the forms is the extent to which arrangement, ruling, spacing, and content have been co-ordinated to emphasize the major results contained.

These forms, of course, are not intended for publication but are prepared for use by persons thoroughly familiar with their contents and purposes. Consequently many things which have required explanation in presenting the forms in this book are commonplace to those who use the forms regularly. This difference in background leads to a general observation of some importance to the budding statistician. Study of the principles and methods of statistics provides a sound basis for engaging in work of the type involved in preparing forms such as these examples, but general knowledge must be supplemented by particular training to produce a practicing statistician.

PROBLEMS

1. What type of classification is employed in each of the following:

⁶ The use of a percentage distribution of separations in the measurement of labor turnover is developed in chapter XII.

A	
COLOR OF HAIR	NUMBER OF STUDENTS IN CLASS
Light	8
Red	3
Brown	7
Black	2

B	
CITY	BANK DEBITS 1938 (Millions of Dollars)
Boston	14,288
New York	168,778
Philadelphia	14,553
etc.	

C	
SIZE OF CITY	NO. OF DWELLING UNITS CONSTRUCTED PER 10,000 POPULATION IN 1940
500,000 and over...	48.6
100,000 to 500,000.	56.8
50,000 to 100,000.	48.5
25,000 to 50,000.	67.4
10,000 to 25,000.	68.7
5,000 to 10,000.	67.3
2,500 to 5,000.	64.4
All urban.....	57.5

D	
YEAR	SHIPMENTS OF FINISHED STEEL BY UNITED STATES STEEL CORPORATION (1,000 net tons)
1937.....	14,098
1938.....	7,316
1939.....	11,707
1940.....	14,976

- From recent issues of any business periodical find three one-way tables, each of which illustrates a classification according to a different kind of characteristic. Copy enough of the table to indicate the kind of classification. Give exact and complete references to the sources used.
- What are the distinguishing characteristics of primary tables and derivative tables?
- Which of the tables of problem 1 are primary and which are derivative?
- In Table 17, page 153, what information is primary and what is derivative?
- The following statistics have been published for the Bell Telephone System:
 "The number of manual service telephones declined from 10,705,118 at the end of 1930 to 9,659,349 at the end of 1931 while the dial service telephones increased in the same period from 4,976,941 to 5,730,645. This is a net decline in number of telephones of 292,065. The average number of telephone calls per day in 1930 was 62,365,000 local and 2,933,000 toll; in 1931 these had declined to 62,205,000 and 2,700,000 respectively, a total decline of 393,000 calls. The miles of wire in underground cable were 50,225,000 at the end of 1930, the miles of aerial cable were 20,785,000. At the end of 1931 there were 52,214,000 miles of underground cable and 21,951,000 miles of aerial cable. There were 5,238,000 miles of open wire at the end of 1930 and 5,074,000 miles a year later."

Present this information in tabular form, taking account of all the points of established practice in table construction. Does your table have unity? Why or why not? What is the degree of complexity? Explain.

7. List the separate classifications present in Figure 24, page 159, and state the characteristic with respect to which each classification is arranged.
8. Study the headings of tables in any year's Supplement to the *Survey of Current Business*. Write a paragraph on the use of table headings based on your findings. Give specific references.
9. a) Consult Table 2, page 1, in any *Statistical Abstract* from 1932 to 1940.
 - (1) Discuss the location of totals. (2) Discuss the arrangement of stub items.
 b) In the *Statistical Abstract* consult either Table 494, page 432, 1939; Table 472, page 408, 1938; Table 464, page 401, 1937; or Table 460, page 401, 1936.
 - (1) Describe the method of classification and degree of complexity.
 - (2) How many different units are there? Name them. Do you think it is justifiable to include all of them in one table? Why or why not?
 - (3) Discuss any desirable or undesirable features in the table
10. Describe in detail the order of arrangement of each of the following tables:

A

DEATHS FROM CHIEF CAUSES IN THE UNITED STATES, 1935

DISEASE	NO. OF DEATHS
Heart	312,333
Cancer	144,065
Nephritis	103,516
Pneumonia	100,279
Accidents	99,967
etc.	

B

BANK CENSUS, 1935

STATE	NO. OF BANKS	NO. OF EMPLOYEES	TOTAL ANNUAL WAGES
Alabama ..	251	2,123	\$ 3,227,296
Arizona ..	39	492	848,587
Arkansas ..	260	1,416	1,905,105
California ..	1,083	19,523	38,675,923
etc.			

C

RETAIL TRADE IN THE UNITED STATES, 1935
APPAREL GROUP

TYPE OF MERCHANDISE	SALES (in millions)
Men's furnishings.....	\$516
Men's clothing.....	144
Family clothing.....	359
Women's ready-to-wear.....	795
Furriers and fur shops.....	60
Millinery stores.....	94
Custom tailors	67
Accessories and other apparel.	110
Shoe stores.....	511

D

VALUE OF PUBLIC BUILDINGS ERECTED IN CITIES IN NEW YORK STATE IN 1936

CITY	VALUE OF THE CONSTRUCTION (in thousands)
Buffalo	\$ 21
Rochester	1,491
Syracuse	1,108
Yonkers	60
Albany	95
Utica	17
etc.	

REFERENCES

Bureau of Agricultural Economics, United States Department of Agriculture, *The Preparation of Statistical Tables, A Handbook*. Washington, D. C., December, 1937.

A statement of the rules for table construction employed by a government bureau.

DAY, EDMUND E., "Standardization of the Construction of Statistical Tables," *Quarterly Publications of the American Statistical Association*, Vol. XVII, No. 129 (March, 1920), pp. 59-66.

A brief but complete statement of the principles of tabulation.

DAY, EDMUND E., *Statistical Analysis*. New York: The Macmillan Co., 1925.

Chapters IV and V contain a detailed discussion of classification of observations.

MUDGETT, BRUCE D., *Statistical Tables and Graphs*. Boston: Houghton Mifflin Co., 1930.

Chapters I, II, and III contain a very lucid explanation of the principles of classification and tabulation.

WALKER, HELEN M., and DUROST, WALTER N., *Statistical Tables, Their Structure and Use*. New York: Bureau of Publications, Teachers College, Columbia University, 1936.

A detailed discussion of the mechanics of table construction and the analysis of tabular material (examples from field of education).

CHAPTER IX

CLASSIFICATION OF LIBRARY SOURCES

THE MEANING OF COLLECTION FROM LIBRARY SOURCES

CHAPTERS IV to VIII have been devoted to the methods of securing data by direct investigation. Chapters IX and X will explain the procedures used in finding data that have already been collected and published. The discussion is introduced at this point because subsequent chapters deal with the steps of analysis which are applicable to data collected either directly or from library sources.

"Library" is used as a general term descriptive of all published sources of business data. Some of the publications which are available to students only in public or school libraries may be kept on file currently by an individual business concern, but the industrial statistician is also likely to be dependent upon libraries for long-time series or other than ordinary data. His method of procedure will not differ materially from that of the student in the search for published data needed in a given problem.

Published sources of business data are for the most part current periodicals or yearbooks. How to become familiar with the contents of these publications presents a very real problem. A reference list of such sources and of their contents is of only temporary value since they are subject to constant changes. New publications may appear and older ones disappear; new series are added, older series are discontinued, and the form of recording is altered. Consequently, in this chapter the emphasis is placed on various classifications of library sources, but no attempt is made to provide a complete list¹ of reference material. The next chapter will deal with the difficulties which may be encountered in securing data from these publications.

METHODS OF CLASSIFYING SOURCES

Published sources of data may be classified in a number of different ways according to the point of view from which a problem is ap-

¹ A selected list of sources is given in Appendix A at the end of the chapter. These sources are numbered consecutively and whenever one of them is mentioned in the text reference is made by number to the detailed description in the appendix.

proached. Five methods of classification are discussed in the succeeding pages: (1) types of data contained, (2) form of publication, (3) frequency of publication, (4) regularity of publication, and (5) publishing agency. These are not all equally important but all of them must be taken into account in acquiring familiarity with sources.

Types of Data Contained

Classification according to type of data is the least important method from a practical point of view, since it is applicable to only a few sources. All source books of business statistics are concerned with economic data, but as a rule they are not confined to a single phase such as production of raw materials, manufacturing, or marketing. *Agricultural Statistics*,² the *Census of Manufactures*,³ and the *Market Data Handbooks*⁴ might be named as representative of these three specific phases of the economic structure. A few other sources of limited scope are: *Statistics of Railways*⁵ (transportation) and *Chain Store Age*⁶ (a specialized type of retail trade). By far the greater number of source books deal with several or all of the functions in our economic system. Examples are: the *Survey of Current Business*,⁷ *Standard Trade and Securities Statistical Bulletin*⁸ and the *Commercial and Financial Chronicle*.⁹

Some of the source books may be classified according to their scope in other respects, for instance, geographically. Many of them contain data for the entire United States; others are confined to a particular state, city, or local area. Still others present world data or data for several different countries. Frequently a single source will cover territorial or governmental subdivisions at various levels. For example, *Agricultural Statistics*² is mainly devoted to a complete compilation of data concerning agriculture in the United States with some information for other countries. In many tables, however, production statistics are subdivided by states, and price and market data by individual cities or regions. This source then is international, national, state, and local in scope, with the major emphasis on the national data.

² Appendix A, No. 23.

³ Appendix A, No. 12.

⁴ Appendix A, No. 6.

⁵ Appendix A, No. 42.

⁶ Appendix A, No. 72.

⁷ Appendix A, No. 1.

⁸ Appendix A, No. 59.

⁹ Appendix A, No. 54.

It can be concluded, therefore, that the majority of sources cover so wide a range of information that they cannot be classified according to specific types of data. This method of classification of source material, although theoretically sound, is not usable in the search for data.

Form of Publication

Statistical Source Books.—Some publications consist almost entirely of statistical tables. Either the index or the table of contents can be used to find the data pertaining to a particular subject. Most publications of this kind come from governmental agencies, although in recent years there has been a great increase in the amount of such work done by private organizations. Examples of the latter are the *Standard Trade and Securities Statistical Bulletin*,¹⁰ *Automobile Facts and Figures*,¹¹ and *A Review of Railway Operations*.¹²

Auxiliary Sources.—Sources which contain data as an auxiliary to other functions are more difficult to use. Data may be scattered through the book or magazine in conjunction with articles to which they apply. This is the case with the *Commercial and Financial Chronicle*,¹³ *Business Week*,¹⁴ and *Commerce Reports*.¹⁵ Other auxiliary sources such as *Dun's Review*¹⁶ and the *Northwestern Miller*¹⁷ group most of their data in one section.

In the great majority of such publications the data appear in tabular form. The tables have proper titles and whatever footnotes are necessary to explain any irregularities of the information. There are, however, a few cases in which valuable data are printed as text material. Careful attention is necessary in order to detect data published in this form and caution must be exercised in using them because necessary explanations may be far removed from the place in the text at which the data are found. It would be advantageous to all concerned if this practice were discontinued, but as long as it persists statisticians must be prepared to search for information appearing in that form. Newspapers avoid the use of tables with some regularity

¹⁰ Appendix A, No. 59.

¹¹ Appendix A, No. 50.

¹² Published annually by the Association of American Railroads, Bureau of Railway Economics, Washington, D. C.

¹³ Appendix A, No. 54.

¹⁴ Appendix A, No. 55.

¹⁵ Published weekly by the United States Department of Commerce, Bureau of Foreign and Domestic Commerce.

¹⁶ Appendix A, No. 57.

¹⁷ Appendix A, No. 66.

because of the difficulty of adjusting them to narrow columns. The following paragraph illustrates the need for a table.

Then, France is a country of handicraftsmen; even after recent and important evolutions, such as the reconstruction of the northern departments and the return of Alsace-Lorraine, it has not decidedly become a country of great industry. Out of 21,721,000 people given to active occupations, only 6,181,000—28%—belong to the industries of transformation. Among those 6,181,000 only 4,027,000—65%—are regular industrial workers; 683,000—11%—are employers, which shows the great number in France of small employers; 1,162,000—18%—work alone independently or are not regularly connected with employers. Out of 4,000,000 of regular workingmen, only 774,000—19%—are employed in factories of more than 500 workers! The conclusion is that France is a country of artisans; the village joiner, the motor-car mechanic, the couturiere even in the great maison de couture, the vine grower, the gardener who raises vegetables or fruits, all belong to that type of workers, and they are undoubtedly the most typical of the French people.¹⁸

Variability of Content.—There is a variation in the form of publishing data which applies to sources devoted entirely to statistical information as well as to those that publish such information incidentally. The most convenient publications to use are those which contain the same series of data in each issue, such as the monthly *Survey of Current Business*,¹⁹ the *Monthly Labor Review*,²⁰ and the *Standard Trade and Securities Statistical Bulletin*.²¹ On the other hand *Crops and Markets*²² and *Steel*²³ present whatever monthly or weekly data are available at the time of publication.

Frequency of Publication

In discussing this classification we will proceed from those sources which appear most frequently to those which have longer intervals between publication dates.

Daily.—Daily papers then are the first source on the list. The financial section of the paper contains information on a variety of subjects. The great virtue of the daily paper is its ability to place data in the hands of its readers quickly. The element of speed tends to

¹⁸ Andre Siegfried, "French Industry and Mass Production," *Harvard Business Review*, Vol. VI, No. 1 (October, 1927), p. 2.

¹⁹ Appendix A, No. 1.

²⁰ Appendix A, No. 16.

²¹ Appendix A, No. 59.

²² Appendix A, No. 24.

²³ Appendix A, No. 62.

reduce accuracy; hence the data found in daily papers are sometimes not reliable and for this reason they should be verified in other sources as soon as possible.

There are a number of daily publications which are valuable sources because they deal with particular subjects. Among these the *Wall Street Journal*,²⁴ the *New York Journal of Commerce*,²⁵ and the *American Metal Market*²⁶ are typical. There are also many daily reports issued by government agencies, such as the daily *Treasury Statement*,²⁷ and daily produce market reports issued by state departments of agriculture.

Weekly.—There is an increasing tendency toward weekly compilation and publication of data to meet the demands of business men for information as nearly current as possible. The demand is further evidence of the extent to which numerical facts have become useful in determining business policy. Accordingly, the Bureau of Labor Statistics now computes its Index of Wholesale Prices weekly. Data such as car loadings, bank debits, and electric power production are available weekly. Among weekly publications the *Commercial and Financial Chronicle*,²⁸ the *Weekly Supplement* to the *Survey of Current Business*²⁹ and *Iron Age*³⁰ are widely used.

Monthly.—Monthly publications also attempt to put information in the hands of users as soon as possible. Sometimes the data for one month are available as early as the 10th of the following month. More commonly the data are one or even two months old before they appear in print. Some important monthly publications are the *Federal Reserve Bulletin*,³¹ the *Survey of Current Business*,²⁹ and bank reports such as the *Business Bulletin* of the Cleveland Trust Company.³²

Annually.—Other valuable sources appear annually. Most important of these are the yearbooks which contain a great amount of basic data with some series running back for long periods. Yearbooks require much preparation, consequently the data may be several months or even a year old before the book is published. Among the valuable

²⁴ Appendix A, No. 58.

²⁵ Journal of Commerce Corporation, New York.

²⁶ American Metal Market Company, New York.

²⁷ United States Treasury Department, published in daily newspapers.

²⁸ Appendix A, No. 54.

²⁹ Appendix A, No. 1.

³⁰ Appendix A, No. 61.

³¹ Appendix A, No. 34.

³² Cleveland Trust Company, Cleveland, Ohio.

yearbooks the *Statistical Abstract*³³ and *Agricultural Statistics*³⁴ may be mentioned. Several newspapers in various parts of the country publish yearly almanacs which contain a large amount of statistical data. These are not usually considered to be authoritative source books but they serve as convenient guides to data which may be found elsewhere.

Longer Intervals.—Examples of sources appearing less frequently are the volumes of the *Census of Population*³⁵ which are published at ten-year intervals, the *Census of Agriculture*³⁶ which has been published along with the *Census of Population* since 1860 and quinquennially since 1925, and the *Census of Manufactures*³⁷ which has been published along with the *Census of Population* since 1850, was also published in 1905 and 1914 and has appeared biennially since 1919.

Special Releases.—Recognizing the necessity of saving time, many of the government bureaus release their more important data as soon as possible. In some instances the data contained in these releases may be preliminary and may subsequently be revised in a regular publication. In other cases the data are not reprinted at any time. The Bureau of Mines has adopted the practice of issuing each chapter of its *Yearbook*³⁸ separately in advance of the complete bound volume. The Bureau of Labor Statistics distributes special processed bulletins on wholesale and retail prices, cost of living, and employment and payrolls, but the greater part of this material is reproduced in subsequent issues of the *Monthly Labor Review*. On the other hand the Bureau of Census releases information concerning the *Census of Manufactures* in both processed and printed form, much of which is never reprinted in the bound volumes. The Bureau of Public Roads of the Department of Agriculture distributes various printed releases concerning state and federal gasoline taxes and automobile registrations and license fees. Some of this information is not reprinted in bound volumes. Knowledge of these various releases must be acquired by experience since they are not always included in check lists of government publications.

³³ Appendix A, No. 8.

³⁴ Appendix A, No. 23.

³⁵ Appendix A, No. 10.

³⁶ Appendix A, No. 11.

³⁷ Appendix A, No. 12.

³⁸ Appendix A, No. 25.

Regularity of Publication

Some published sources appear at regular intervals; others appear irregularly.

Regular.—The question of regularity is important from several points of view. Business men have learned to expect weekly publications to be delivered in a certain mail. They delay decisions until the arrival of the latest statistical report. Absolute regularity of publication is required to meet this demand. Statisticians depend upon regular publications for current data in carrying on their research work. The business community in general expects to receive regular information from weekly or monthly publications. In other cases no exact date of publication is observed but publication is certain each week, or month, or quarter, or year. Regularity is, of course, a great virtue in source material and the great majority of publications possess it.

Irregular.—There are other publications which appear irregularly although the data which they contain are collected quite regularly. The *Census of Manufactures* contains data that are collected biennially but are published whenever the Department of Commerce is able to prepare the data and funds are available to meet the cost. Only those irregular publications that are already in print can be included in the plan of an investigation. On the other hand regular publications which are scheduled to appear while an investigation is in progress can be included. An example will perhaps clarify this distinction. As this is being written parts of the results of the 1940 *Census of Population* have been released. It would not be possible, however, to plan an investigation requiring the use of the complete *Census* or any unpublished part of it because there is no way of knowing when the needed data will be published.

Special Studies.—While the most valuable sources are those which are published at regular intervals and which consequently keep the information up-to-date, there are many special studies that appear from time to time containing data which are available in no other sources. These studies are reports of special researches. Usually they are models of the application of statistical methods in practical work. They are consequently valuable references for data, methods of analysis, and form of presentation. Examples of this type of work are the *Cost of Living Studies* of the Department of Labor made in 1918 and in 1936.⁸⁹

⁸⁹ Appendix A, No. 22.

There are also important non-government publications of this type among which the *Retail Clothing Survey* made by Northwestern University⁴⁰ and the *Study of Income in the United States, 1909-19*, by the National Bureau of Economic Research⁴¹ are excellent examples.

Publishing Agency

This classification of sources has probably the greatest practical value in library research. The publications are usually catalogued on this basis in the libraries. In discussing this classification it will be well to recall that only those sources that are related to the field of business are included, no attempt being made to include sources of statistical data in other fields.

United States Government.—The most important publishing agency is the federal government, chiefly through the executive branch which includes all the departments and independent offices.⁴² An outline of its plan of organization is given in Figure 32. Some of the departments and certain bureaus and offices in particular produce a tremendous amount of statistical data, whereas others by the very nature of their work produce none. In searching for data, one quickly becomes familiar with the Bureau of Foreign and Domestic Commerce and the Bureau of the Census in the Department of Commerce; the Bureau of Labor Statistics in the Department of Labor; the Bureau of Agricultural Economics in the Department of Agriculture; the Bureau of Internal Revenue in the Department of the Treasury; the Bureau of Mines in the Department of the Interior; the Interstate Commerce Commission; the Federal Trade Commission; and the Board of Governors of the Federal Reserve System.

The list of sources given in Appendix A at the end of the chapter is arranged according to publishing agency and includes the most important statistical publications of these bureaus and other offices.

In addition to the regular government publications which are usually issued as periodicals or yearbooks, there are often useful data in the annual reports of department, bureau, and division heads. Several of

⁴⁰ *Costs, Merchandising Practices, Advertising and Sales in the Retail Distribution of Clothing*, Vol I-VI, 1921. *Selling Expenses and Their Control in the Retail Distribution of Clothing*, Vol VII, 1922. New York: Prentice-Hall, Inc.

⁴¹ Volumes I and II of the Publications of the National Bureau of Economic Research, New York

⁴² *The United States Government Manual*, published three times yearly by the Office of Government Reports, Washington, D. C., gives complete and up-to-date information on the organization and activity of all subdivisions of the federal government. The outline in Figure 32 is reproduced from this source.

these are listed in Appendix A. Another is the *Annual Report of the Commissioner of Immigration* (Department of Labor), which gives data concerning immigrants and emigrants in more detail than can be found in any other publication. Some annual reports are available only as numbered documents of the Congress to which they were submitted, but others are published separately.

Finally, the special investigations made for congressional committees should be mentioned. These are usually detailed studies of a particular subject and as such are unique sources. They are likewise usually published as Congressional Documents. Excellent examples are the Marine Insurance Investigation of 1920⁴³ and the Chain Store Investigation of 1929-33.⁴⁴

State and Municipal Government.—In many cases the most prolific sources for information concerning the individual states are the publications of the federal government that have already been mentioned. In addition there are many publications by the state governments themselves. The latter vary so much from state to state that an attempt to list them would not be feasible. Some states are far in advance of others in furnishing statistical information to their citizens. A few have begun the publication of yearbooks similar in plan to the *Statistical Abstract* of the United States. A list of sources of market data available for the various states is given in *Market Research Sources*.⁴⁵ This list can be supplemented by consulting the library card catalog under the individual states. Information is likely to be found in the publications of the state departments of Agriculture, Banking and Insurance, Labor, and Highways. The publications of the Land Grant Colleges and other state institutions also contain valuable special data.

Very few source books are published by municipalities, but considerable local information is available in federal and state publications. For example, the *Census of Manufactures*⁴⁶ includes data for individual cities; likewise the *Monthly Labor Review*⁴⁷ gives a retail food price index monthly for 51 cities; and the *Industrial Bulletin*⁴⁸

⁴³ S. S. Huebner, *Report of Status of Marine Insurance in the United States, and Report on Legislative Obstructions to the Development of Marine Insurance in the United States*.

⁴⁴ Investigation for the Senate Committee on the Judiciary by the Federal Trade Commission. Published in several parts in 1933 as Numbered Documents of the 72nd Congress.

⁴⁵ Appendix A, No. 7.

⁴⁶ Appendix A, No. 12.

⁴⁷ Appendix A, No. 16.

⁴⁸ New York State Department of Labor.

provides monthly data on employment and payrolls for various cities in New York State.

Non-Government.—State and local information is also available through the publications of certain semi-public organizations. The best examples are the monthly reviews of business issued by the 12 Federal Reserve Banks⁴⁹ and reports of statistical studies by the research bureaus of universities.

In addition there are many private agencies which publish statistical data for general use. In a majority of cases the data are collected for the use of an interested group such as the members of a trade association or the subscribers to a service, but are made generally available through magazines and trade papers. There are other cases in which data are collected and published merely to increase the value of a magazine to the reading public. In any event the cost of making the data available must be borne by the subscribers to the publication. Historically, private agencies preceded the government in supplying current data to the public. Among the pioneers in this field were *Dun's Review*, *Bradstreet's Review*, *The Commercial and Financial Chronicle*, and *Babson's Service*.

The private agencies compiling and publishing statistical information may be classified as follows: trade, industrial, and financial associations; financial magazines; statistical services; and trade and industrial magazines. The outline of sources used in Appendix A conforms to this one. The examples given there are some of the most commonly used non-government sources. The list has purposely been confined to only a few of the multitude of publications which might have been included. Many of those omitted contain data of value; hence the need for gradually expanding one's knowledge of them as progress is made in the use of source material.

Foreign and International.—While the agencies already named provide most of the data needed for statistical work, there are occasions which call for the use of information from foreign countries or for world data. A list of sources published in foreign countries as well as publications containing world data can be found in *The Economists' Handbook—A Manual of Statistical Sources*.⁵⁰

⁴⁹ Appendix A, No. 36. These are actually private organizations but because of their close integration with the Federal Reserve system their bulletins have been listed as government publications.

⁵⁰ Verwey and Renooiz, Amsterdam, 1934.

Summary

The use of published sources of business data requires a knowledge of government and non-government publishing agencies and the titles of their publications, as well as a knowledge of the form, frequency, and regularity of publication and the type of data contained. The classification according to publishing agency is the most generally usable one; hence it has been employed in Appendix A.

It is necessary to remember, however, that source material is continually changing with respect to all of these classifications. That is, the type of data may be changed by the addition of new series and the elimination of old ones, or by changing the titles of series and the data included in them. Likewise the form of publication may change. Data formerly scattered throughout the publication may be brought together in a statistical appendix or published separately. On the other hand statistical compendiums may be abandoned. Again the frequency of publication changes when a weekly publication becomes monthly or the reverse; when an annual is supplemented by a monthly and/or weekly or when a weekly or monthly issue begins publication of an annual supplement. The regularity of publication also undergoes changes. Sources which have appeared regularly for years may be discontinued entirely⁵¹ or subsequently may appear irregularly. Finally, changes occur in publishing agency and in titles of publications. For example, the material formerly found in *Bradstreet's Review* is now found in *Dun's Review*; the Bureau of Mines of the Department of Interior was for several years in the Department of Commerce; the former *Commerce Yearbook, Volume II, Foreign Commerce* is now the *Foreign Commerce Yearbook*.

In the light of these changing conditions it may readily be understood that any list such as that given in Appendix A loses its accuracy after a few years. It is therefore necessary for users of published source material to keep abreast of current changes as they occur. The list given in the appendix provides an adequate nucleus which can be kept up-to-date by noting additions and changes from time to time.

⁵¹ One of the most useful references, *The Annalist*, heretofore published weekly by the New York Times Co., was discontinued in October, 1940. Although many of the series of data carried by this publication are no longer available currently, the volumes for earlier years contain valuable data.

APPENDIX A

SELECTED SOURCES LISTED ACCORDING TO PUBLISHING AGENCY, TITLE,
FREQUENCY OF PUBLICATION, AND CONTENTS
(REVISED TO OCTOBER, 1940)

UNITED STATES GOVERNMENT SOURCES

Department of Commerce—Bureau of Foreign and Domestic Commerce

1. *Survey of Current Business* (monthly and weekly, with occasional yearly supplements)

The monthly issues present data for the United States concerning business indexes, commodity prices, construction and real estate, domestic and foreign trade, employment, finance, transportation and communication, and statistics of industry in 12 general subdivisions; also some Canadian data. Brief summary statements and tables are given concerning each, followed by monthly statistics that cover the preceding 13 months.

The weekly pamphlet brings some of the monthly series up-to-date in advance of the monthly issue, and gives weekly data for a few important series.

The supplements to date have been issued for the years 1931, 1932, 1936, 1938, and 1940. Each covers a number of years, and taken together (except for subsequent revisions) they furnish continuous monthly data, including monthly averages for every year since each series has become available. Full notes are appended explaining the sources and methods of construction of each series.

2. *Monthly Summary of Foreign Commerce of the United States* (monthly)

Gives dollar value and quantity of all goods exported from and imported into the United States including gold and silver. Includes a detailed classification of both exports and imports by articles and by customs districts. Since each month's issue contains data for that month and the calendar year to date the December issue contains the total for the year.

3. *Domestic Commerce* (weekly)

Presents digests of important studies both government and non-government, summaries of federal bills, laws, and court decisions, also statements of recent publications by the various government agencies, a list of recent publications dealing with domestic commerce, and some data regarding changes in wholesale and retail trade along specific lines.

4. *Foreign Commerce Yearbook* (annually)

Purpose is "to provide in a single volume all the important basic statistical material essential for a comprehension of current economic developments in foreign countries." Part I gives data for each country separately, total trade and trade with the United States being given by

specific commodities; Part II gives comparative world statistics on population, production in agriculture and industry, and trade.

5. *Foreign Commerce and Navigation of the United States* (annually)

Detailed tables of specific items of export and import by countries; also number and tonnage and ports of arrival and clearance of American and foreign vessels.

6. *Consumer Market Data Handbook, 1939* (intervals of 3 or 4 years)

A nation-wide survey of the markets for consumer goods presenting "on a comparable basis for all counties, and in most cases for all urban communities, just how much consumers spent in retail stores, and in service, amusement and hotel establishments; what wholesale business amounted to; how many of the consumers' homes had telephones and electric meters; how many persons made out an income tax return; how many passenger automobiles were registered; what the relief load amounted to; and other factors indicative of the relative importance of each market."

An *Industrial Market Data Handbook* was also published in 1939, giving data regarding the industries in every county in the United States.

7. *Market Research Sources* (biennially)

Complete and thoroughly cross-referenced lists of all government and non-government sources relating to problems of domestic marketing.

Department of Commerce—Bureau of Census

8. *Statistical Abstract of the United States* (annually) Prior to 1938 published by Bureau of Foreign and Domestic Commerce.

"A digest of data collected by all statistical agencies of the national government," as well as by some states and private agencies. It consists entirely of summary tables and time series, chiefly annual data, with notes defining the scope and terms used in each, and referring to the original sources.

9. *Abstract of the Census* (decennially)

A selection of the most essential statistics collected on all subjects at each census, in one volume. Data are given by subjects, states, and cities, and some by counties and smaller civil subdivisions. Some data are included covering outlying territories and possessions of the United States.

10. *Census of Population* (decennially)

Data are classified by states, counties, cities or villages, and minor civil subdivisions. The subjects included are: color, race, nativity, parentage, origin of foreign born, sex, marital condition, age, urban and rural distribution, citizenship, school attendance, and literacy. Separate volumes give data on occupations and families, and sometimes unemployment. Reports on special groups or subjects are published at irregular intervals

in intercensal years, such as *Religious Bodies, Benevolent Institutions, The Blind and the Deaf, Negroes in the United States*, etc. A series of *Mortality Statistics* is published annually.

11. *Census of Agriculture* (quinquennially)

The year which coincides with a decennial census affords somewhat more detailed information than the intervening non-census year. In both, data are given by states and counties, for the number of farms, color and tenure of farm operator, uses of farm land, value of land and buildings, acreage, production and value of specified crops, and value of livestock by principal classes and age groups. In the decennial years, classifications are made also by minor civil subdivisions, and special reports such as irrigation are included.

12. *Census of Manufactures* (biennially)

The reports for years which coincide with a decennial census are given in greater detail than those taken in the intervening non-census years. Mines and quarries have been covered only in the decennial years (see 1935 report under *Census of Distribution*). All reports give data concerning number of establishments; number of wage earners and salaried employees; amount of wages and salaries paid; cost of materials, fuel, and power; value of products; and value added by manufacture. Classifications are by industry groups, by states, by cities, and in some years, notably 1929, by industrial areas.

13. *Census of Distribution, and Business Censuses*

(1) *Census of Distribution, 1929*

This first attempt to gather nation-wide business statistics was a part of the fifteenth census. It included retail trade, wholesale trade, distribution of manufacturers' sales, contract construction, and hotels.

(2) *Census of American Business, 1933*

This covered the same field as the 1929 *Census of Distribution*, with the addition of services and places of amusement. Both afford data on number of establishments, net sales or receipts, personnel, and payroll. They are classified by field and kind of business, and by states, with some basic data for cities and counties.

(3) *Census of Business, 1935*

This is more comprehensive than either of the preceding censuses. Subjects added are: transportation and warehousing, tourist camps, radio broadcasting and advertising agencies, banking and finance, insurance, mines and quarries.

(4) *Census of Business, 1939* (part of sixteenth census)

Will be practically the same as 1935.

14. *Financial Statistics of Cities* (annually)

Shows the financial transactions of cities having a population of over 100,000, including taxes, indebtedness, specified assets, government costs and receipts.

15. *Financial Statistics of State and Local Governments* (published decennially several years subsequent to census)

"Statistics relating to revenue receipts, governmental cost payments, public debt, and assessed valuations and tax levies, for all divisions 'of government.'" The classifications are by states, counties, cities, towns, villages, boroughs, school districts, townships, and other civil divisions. *Financial Statistics of States* is also published annually when funds permit, but there was none between 1931 and 1937.

Department of Labor—Bureau of Labor Statistics16. *Monthly Labor Review* (monthly)

Contains brief reports and complete statistical data on all matters handled by the department. Each issue contains sections on industrial disputes, wages and hours of labor, employment and payrolls, wholesale and retail prices, and cost of living, and usually other sections on labor legislation, industrial accidents, etc. There are always a few special articles on timely subjects concerning labor, and a list of recent publications by the department.

17. *Wholesale Prices* (monthly)

Monthly index numbers of wholesale commodity prices by groups and subgroups. Each issue gives the group and subgroup indexes and detailed indexes for certain groups. Comparisons are given with the same month in previous years, and with foreign countries. The December issue gives indexes for the 12 months of the year for the entire series of more than 800 commodities.

18. *Retail Prices* (monthly)

Index numbers of retail prices of food, coal, electricity, gas, and other consumers' goods. Food data are given every month; coal, electricity and gas at frequent intervals; and other commodities less frequently.

19. *Changes in the Cost of Living* (quarterly)

Index numbers of changes in the cost of living, divided into 6 groups: food; clothing; rent; fuel, electricity, and ice; house furnishings; and miscellaneous, for 33 cities.

20. *Employment and Payrolls* (monthly)

Index numbers of employment, payrolls, hours worked, and weekly earnings for all manufacturing and non-manufacturing industries. It includes data for employment and payrolls in the regular executive departments of the federal government and on emergency work.

21. *Labor Information Bulletin* (monthly)

Brief summary of labor conditions. Hours of work, wages, cost of living, employment and payrolls, wholesale prices, retail food prices, industrial production and trade, agriculture, and government employment and relief for the month.

22. *Numbered Bulletins* (irregular, several each year)

"Each bulletin contains matter devoted to one of a series of general subjects, those subjects being Wholesale Prices, Retail Prices and Cost of Living, Wages and Hours of Labor, Employment and Unemployment," and many other subjects of interest to labor, but not of a statistical nature. Bulletin No. 661 gives a selected list of these bulletins as of 1938, and the most recent ones are listed on the back cover of each *Monthly Labor Review*. In recent years the tendency appears to be not to issue these bulletins on subjects covered by the regular monthly pamphlets listed above, so that the majority of the current bulletins deal with Wages and Hours of Labor in specific industries. One of the most important of the series is No. 357, *Cost of Living in the United States*, published in 1924 and giving a complete statistical report of the first extensive cost-of-living study, made in 1918-19. The more recent study made in co-operation with the Works Progress Administration is being reported in a series of bulletins beginning in 1936 under the general title "Studies of Consumer Purchases."

On June 30, 1939 a special pamphlet (unnumbered) entitled *Publications of the Department of Labor* was issued. This contains a complete list of all the publications of the various bureaus of the Department of Labor since it was organized. Of particular value are the detailed descriptions of changes of form and content which were made during the entire period in the various series published by the department.

Department of Agriculture

23. *Agricultural Statistics* (annually)

Prior to 1936 was included in the *Yearbook of Agriculture*. Presents summary tables of all statistical data appearing in periodicals of the department, in great detail, usually covering a series of years. These include data on all United States crops, livestock, poultry and dairy products; farm business; foreign trade in agricultural products; and some data on world production.

24. *Crops and Markets* (monthly)

Each issue gives complete detailed reports, estimates, and forecasts on United States crops and other farm products, the items included varying with the seasons. Sectional data are given, and comparison with preceding years. Prices, wages, labor supply, stockyards, and market reports, and some items of export and import are included.

Department of the Interior—Bureau of Mines**25. *Minerals Yearbook* (annually)**

Gives a general survey of mineral production in the United States and the world, and separate chapters dealing with each metal and non-metal. It contains the most recent compilation of statistical data on the more important minerals: coal, natural gas, petroleum, stone, gold, silver, copper, lead, and zinc, etc.

26. Several **weekly and monthly** bulletins on production and distribution of anthracite coal, bituminous coal, and coke.

War Department**27. *Report of the Chief of Engineers, U. S. Army, Part 2, Commercial Statistics* (annually)**

A complete review of water-borne commerce of the United States both domestic and foreign, freight and passenger, subdivided by grand divisions, by ports, and by commodities. All data are annual for calendar years.

Post Office Department**28. *Annual Report of the Postmaster General* (annually)**

Contains detailed statistical analysis of all receipts and expenditures of the department, for the fiscal year ending June 30; the number of post offices and employees; mail carried by each type of carrier; and money order transactions.

Treasury Department**29. *Annual Report of the Secretary of the Treasury on the State of the Finances* (annually)**

Statistical data for the fiscal year ending June 30 on receipts, expenditures, deficit, public debt, and monetary developments. The "exhibits" on public debt contain statements of all outstanding obligations (bonds, treasury notes, treasury bills, treasury savings certificates, and currency) issued by the United States government. Also list of securities owned by United States government, and statement of assets and liabilities of government corporations and credit agencies of the United States.

30. *Annual Report of Comptroller of Currency* (annually)

Report for the fiscal year ending October 31 covering in great detail all banking operations in the United States and money in circulation. Statements are included of Reconstruction Finance Corporation, Farm Credit Administration, Federal Home Loan Bank System, Federal Deposit Insurance Corporation, Pacific National Agricultural Credit Corporation, and United States Postal Savings System.

31. *Combined Statement of Receipts and Expenditures, Balances, etc., of the United States* (annually)

Very detailed statement of receipts and expenditures of each department and independent office.

Treasury Department—Bureau of Internal Revenue

32. *Statistics of Income* (published annually, about 2 years late)

Detailed data on income tax returns by individuals, partnerships, and corporations, estate tax returns, and gift tax returns for the United States and individual states. Data for counties, cities, and towns are available in separate mimeographed bulletins. Beginning with 1934 corporation tax returns are published separately as Part II.

33. *Annual Report of Commissioner of Internal Revenue* (annually)

Detailed statistical report for the fiscal year ending June 30 on all tax revenue of the United States, including income taxes and all other miscellaneous taxes. A few of the tables give monthly data and comparison with previous years.

Board of Governors of Federal Reserve System

34. *Federal Reserve Bulletin* (monthly)

The only official statement by the Board concerning the operations of Federal Reserve banks and member banks. Monthly or weekly data are given concerning financial, industrial, and commercial statistics in the United States; international financial statistics; and several indexes constructed by the Division of Research and Statistics of the Federal Reserve Board on industrial production, construction, employment and payrolls, freight car loadings, and department-store sales. Summaries and discussion of current financial events, legislation, etc., appear in each issue.

35. *Annual Report of the Board of Governors of the Federal Reserve System* (annually)

Data similar to those in the monthly issues, but given for a series of years, some dating back to 1914.

36. *Monthly Publications of Federal Reserve Districts* (monthly)

The Federal Reserve Bank of each of the 12 districts publishes a monthly bulletin summarizing business conditions in that district.

Federal Home Loan Bank Board

37. *Federal Home Loan Bank Review* (monthly)

Contains data concerning housing and building conditions including building permits, building costs, mortgages, building and loan association activity, government housing activity. It includes the monthly operating

statement of the Home Owners Loan Corporation and the financial statement of the Home Loan Banks. Important special articles on housing appear in each issue.

38. *Annual Report of Home Loan Bank Board* (annually)

Contains data for recent years similar to those in monthly issues.

Federal Power Commission

39. *National Electric Rate Book and State Rate Books* (periodic intervals)

40. *Monthly Bulletin* (monthly, and an annual summary)

Production of electric energy in the United States, sources of energy by states, average daily production by public utility plants.

Federal Communications Commission

41. *Operating Data from Monthly Reports of (a) Telephone; (b) Telegraph Carriers* (monthly)

- a) A report of detailed operating revenues, operating expenses, income items and changes in capital items, by regions for telephone carriers giving the current month and cumulative totals for the year to date.
- b) A report of detailed revenue, expenses, and income of individual telegraph carriers giving the current month and cumulative totals for the year to date.

Interstate Commerce Commission—Bureau of Statistics

42. *Statistics of Railways in the United States* (annually)

Summary statements concerning equipment, employees, revenues, expenses, and other data for all steam railways in the United States usually classified by districts. There are also separate reports from each company.

43. *Annual Report of the Interstate Commerce Commission* (annually)

Contains a statistical appendix giving data on railway development for a preceding period of years. Also contains miscellaneous summaries of data on operating revenue, expense and income, operating ratios, employment, and car loadings.

44. *Freight Commodity Statistics Class I Steam Railroads in the United States* (annually)

Includes annual data on car loadings by districts and groups of commodities for the preceding ten years, also quarterly data by districts and commodity groups for the latest available year. The data are also broken down into individual commodities carried by individual railroads.

A quarterly supplement to this report, having the same title and giving the car loadings for the most recent quarter classified by districts and by individual commodities is also published.

45. *Wage Statistics of Class I Steam Railways in the United States* (monthly)

A complete statement by occupations of the number employed, time worked, and wages received, with summaries.

46. *Statistics of Class I Motor Carriers* (annually)

Data regarding motor transportation of property and passengers.

NON-GOVERNMENT SOURCES

Financial, Trade, and Industrial Associations

47. Reports of National Industrial Conference Board by National Industrial Conference Board, Inc., New York

a) *The Economic Record* (semi-monthly)

Data on wages, earnings, hours, and employment by individual industries; also cost of living. All data except retail food prices are collected by the Conference Board, and are independent of similar series published by the Bureau of Labor Statistics.

b) *The Management Record* (monthly)

Data similar to those in *Economic Record* with articles of interest to employers.

c) Special studies (irregularly) as supplements to *The Economic Record*.

48. *Annual Statistical Report of American Iron and Steel Institute* (annually) New York

Data concerning all iron and steel products, classified by types and by states for a period of years. The report includes also data on prices, foreign trade, production in other countries, and some information on allied industries, as coal and coke.

49. *Electrical Research Statistics* (monthly) by the Edison Electric Institute, New York

A single sheet giving classified data of production, consumption, and sales of electric power. A similar sheet is also issued weekly, and a more comprehensive annual bulletin. This series supersedes similar bulletins published until 1937 by the National Electric Light Association, New York.

50. *Automobile Facts and Figures* (annually) by Automobile Manufacturers Association, New York

Devoted exclusively to data related to the automobile industry; includes production, sales, registration, taxation, financing, exports, truck transportation, used car sales, and allied data.

51. *Statistical Bulletin* (annually) by American Petroleum Institute, New York

A complete collection of data relative to the petroleum industry including production, consumption, imports and exports, and stocks on hand for the various products. The data are given monthly for the last two years and annually back to 1918.

Monthly supplements of the *Statistical Bulletin* give current figures comparable with those in the annual issue. Additional current data dealing mainly with crude oil production by producing areas are supplied weekly.

52. *Exchange* (monthly) by New York Stock Exchange, New York

Supersedes the *New York Stock Exchange Bulletin*, giving summary data of the activities of the New York Stock Exchange, number and volume of sales, etc.

53. *Monthly Survey of Life Insurance Sales in the United States* (monthly) by Life Insurance Sales Research Bureau, Hartford, Conn.

A report of new ordinary insurance written for the current month and for the year to date, subdivided by states and regions. In 1937 a special report was published giving revised sales figures monthly from 1930 to 1936.

Financial Magazines and Papers

54. *The Commercial and Financial Chronicle* (weekly) by Wm. B. Dana Co., New York

Gives stock and bond quotations on the various exchanges for the preceding week, banking and financial data currently reported, corporation balance sheets and statements, general industrial, trade and commodity data, news, and comments. Difficult to use because of variable content from week to week, but a valuable source for a wide variety of data.

55. *Business Week* (weekly) by McGraw-Hill Publishing Co., Inc., New York

A page entitled "Figures of the Week" contains data on production, trade, prices, finance, and banking.

56. *Barron's* (weekly) by Barron's Publishing Co., Inc., New York

Material very similar to the *Commercial and Financial Chronicle*, but presented in somewhat more popular style. Features several original indexes, barometers, etc.

57. *Dun's Review* (monthly) by Dun & Bradstreet, Inc., New York

General analysis of business conditions including regional indexes. The original source for data on business failures and indexes of wholesale commodity prices.

Each month *Dun's Statistical Review* is published as a supplement to the regular magazine. This supplement contains more detailed data on the subjects included in the magazine.

58. *Wall Street Journal* (daily, except Sundays and holidays) by Dow-Jones & Co., Inc., New York

Current events of economic and financial interest in United States and world. The previous day's quotations on stocks and bonds on exchanges throughout the country and in foreign countries, as well as commodity information are given. Indexes of stock, bond, and commodity prices, dividend payments, and industrial data are included.

Statistical Services

59. *Standard Trade and Securities, Statistical Bulletin* (annually and monthly) by Standard Statistics Co., Inc., New York

A complete record of monthly data concerning business and financial operations running back as far as the data are available. This is one of the most valuable reference books in print. The series are kept current in monthly supplements which may be bound with the most recent annual volume.

60. *Moody's Manuals of Investment* (annually) by Moody's Investors Service, New York

Contains financial statements of several thousand corporations both domestic and foreign, including a brief history, balance sheet and income statement of each corporation and a record of securities in the hands of the public. There are five volumes published each year—industrials, railroads, public utilities, banks and finance, government and municipals.

Trade and Industrial Magazines

(The titles of these magazines give sufficient indication of the type of data contained, consequently the descriptions have been omitted)

61. *Iron Age* (weekly) by Chilton Co., Philadelphia, Pa.
62. *Steel* (weekly) by Penton Publishing Co., Cleveland, Ohio
63. *Metal Statistics* (annually) by American Metal Market Co., New York
64. *India Rubber World* (monthly) by Bill Brothers Publishing Co., New York
65. *Textile World* (monthly) by McGraw-Hill Publishing Co., Inc., New York
66. *Northwestern Miller* (weekly) by The Miller Publishing Co., Minneapolis, Minn.
67. *Automotive Industries* (weekly) by Chilton Co., Philadelphia, Pa.
68. *Railway Age* (weekly) by Simmons-Boardman Publishing Co., New York
69. *Marine Engineering and Shipping Review* (monthly) by Simmons-Boardman Publishing Co., New York

70. *Engineering and Mining Journal* (monthly) by McGraw-Hill Publishing Co., Inc., New York
71. *Printers' Ink* (weekly) by Printers' Ink Publishing Co., New York
72. *Chain Store Age* (monthly) by Chain Store Publishing Co., New York

PROBLEMS

1. A young stockbroker interested in general business conditions is planning a small library of statistical source material. The following list has been selected as adequate: *World Almanac*, current year; subscription to *Monthly Summary of Foreign Commerce*; subscription to *Commercial and Financial Chronicle*; subscription to *Business Week*; *Statistical Abstract of the United States*, most recent volume; Vol. I of *Population, Sixteenth Census*.
 - a) Which of the foregoing would you retain?
 - b) Name four others that should be included.
 - c) Give reasons for your choice in (a) and (b).
2. Name the publications that correspond to the following descriptions:
 - a) Published monthly, by a government agency, containing some textual material and about 50 pages of tables that are practically identical in form from month to month, chiefly on the subject of finance.
 - b) A 4-page leaflet published weekly by a government agency, containing certain indexes and other current weekly data in every issue; and also a number of series of monthly data, some of which appear in one issue and some in another, during each month.
 - c) A group of large volumes published annually by a private company, each volume of which contains complete information concerning individual corporations of a certain type.
 - d) A monthly government publication dealing solely with exports and imports, the December issue of each year constituting a summary of that year's data.
 - e) A volume published by a private statistical concern, containing long series of monthly data and index numbers on every phase of business, the series being kept up-to-date by the addition of current supplements.
 - f) A series of volumes, published at intervals of an irregular number of years and under slightly different titles, by a government agency, each issue containing the most complete data available in the United States on trade and various aspects of business other than industrial production.
 - g) A weekly periodical, non-government, each issue of which contains current data on steel prices, with much more complete data on production, shipments, etc., in a large special number issued during January of each year.

3. From the following list, describe each publication according to the five methods of classification named in chapter IX (instructor will assign one or more to each student according to the time available): (a) *Monthly Labor Review*; (b) *Minerals Year Book*; (c) *Survey of Current Business*; (d) *Abstract of the Census*; (e) *Census of Business*; (f) *Statistical Abstract*; (g) *Monthly Summary of Foreign Commerce*; (h) *Moody's Manuals of Investments*; (i) *Commercial and Financial Chronicle*.
4. (Class exercise.) Name a source in which you think each of the following sets of data would be available. Explain your choice in each case.
 - a) The number of tons of pig iron produced in the United States monthly from 1929 to 1936 inclusive.
 - b) The number of employees on the payrolls of manufacturing concerns in the United States in 1934, 1935, and 1936.
 - c) The number of gallons of gasoline consumed monthly in the United States during the first six months of last year.
 - d) The number of automobiles produced in the United States in 1935.
 - e) The amount of sales by chain grocery stores in the state of New York in 1939.
 - f) The value of agricultural products exported by the United States during the most recent month.
 - g) The number of freight car loadings of grain and grain products shipped in the United States during the year before last.
 - h) The index of department-store stocks for the United States, for the most recent month.
5. Give exact reference to a publication (not mentioned in the text) containing numerical data not in tabular form.
6. List publications (not mentioned in this section of the text) illustrating each subheading under the classification "Frequency of Publication."
7. Give exact reference to a special statistical study (not mentioned in the text).

CHAPTER X

THE USE OF LIBRARY SOURCES

INTRODUCTION

A SUPERFICIAL consideration of the matter might easily lead one to expect that the entire task of collecting data from library sources consists in copying a quickly discovered list of figures from a book readily supplied by a library attendant. This is not what usually happens. Only in highly specialized libraries will an attendant be found who is trained in the intricacies of source material. In most cases the library staff will not be able to render any greater service than that of obtaining books and magazines from the stacks on request.

Efficiency in collecting data from libraries comes only with long practice. It is a case primarily of learning to know what data to expect in different sources. While the beginner has no choice but to use what might be called the "shotgun" method, that is, search until the desired data happen to be found, a seasoned investigator uses a process of elimination based on his previous experience to narrow his search to two or three likely sources. By contrast this might be called the "rifle" method. If his selection has been accurate very little time will be required to find the data, or to obtain a guide as to where they may be found, or to discover that they are not available. In passing from the "shotgun" to the "rifle" method, there are two major things to be considered: (1) how to find a good source and (2) how to use it after it has been found.

FINDING A GOOD SOURCE

The purpose of this section is to set up a sequence of steps which can be generally employed in searching for a desired set of data. The process is one of successive elimination, but some guidance in the order of procedure will facilitate the work. There are usually two stages in the search, finding data on the general subject and finding a particular set of data. There is no way of knowing in advance at what point the search should be concentrated on specific information. That must be determined in individual cases according to the circumstances.

Steps

The following steps are suggested in making a search of library sources.

Step 1.—There are several standard reference sources which should be consulted for information on the desired subject. These are: *Statistical Abstract of the United States*,¹ *Agricultural Statistics*,² *Survey of Current Business*,³ *Monthly Labor Review*,⁴ *Federal Reserve Bulletin*,⁵ *Standard Trade and Securities Statistical Bulletin*.⁶ Look in the indexes of these publications for the subject of the search. If the particular data can be obtained from one or several of them the search is ended.

Step 2.—If the desired data cannot be found in these sources, study the titles, headnotes, footnotes, and references of tables on the general subject to discover the original sources which may contain more detail. Study these detailed sources in turn for references to collateral sources.

Step 3.—If steps 1 and 2 have not led directly to the publication containing the precise information required, it is time to consult a bibliography of source material. The current edition of *Market Research Sources*⁷ provides the most useful guide for any subject related to domestic marketing. It contains a full "finding guide" of subjects followed by a list of government and non-government publications classified according to publishing agency. Books and yearbooks are included as well as periodicals.

If current data are desired, it is quite likely that their origin can be traced through the use of another publication of the United States Department of Commerce entitled *Sources of Current Trade Statistics*.⁸ This book is arranged in ready reference form so that the source of a particular series of data can be found through a finding index in the first part of the book and a list of sources in the second part. Neither the finding index nor the list of references includes any annual publications or statistical compendiums. For example, the *Statistical Abstract* and the *Standard Trade and Securities Statistical Bulletin* are not mentioned. This guide does, however, include some references on

¹ Appendix A, No. 8.

² Appendix A, No. 23.

³ Appendix A, No. 1.

⁴ Appendix A, No. 16.

⁵ Appendix A, No. 34.

⁶ Appendix A, No. 39.

⁷ Appendix A, No. 7.

⁸ Latest edition to date, June, 1937.

foreign trade which is one of the few subjects not covered by *Market Research Sources*. Data expressed in index number form can often be located by referring to *An Index to Business Indices*.⁹ This book contains a finding index that is convenient to use in locating the detailed descriptions of indexes appearing in the second part of the book. These descriptions include the names of the source or sources in which the desired index can be obtained.

Step 4.—At this point the card catalogue of the library should be consulted if the data have not been found. The cards are classified by author, title, and subject. Look up the subject concerning which you want to get the data. You will probably find references to non-government publications. Select those which are likely to contain data and investigate them. If the data are still elusive the next reference should be to the government list of publications. These are sometimes not included in the main subject catalogue of the library but are listed separately under "United States." The classification is by departments, bureaus, commissions, and offices. The publications most likely to yield results are listed in Appendix A.

Step 5.—Each month the Government Printing Office issues the *Monthly Catalogue of United States Public Documents* which includes all publications during that month. Several monthly catalogues should be examined to discover any recent publications on the subject of the search. This "check list" is classified by departments, bureaus, etc.

Step 6.—If access to the stacks of the library is possible, the search should be continued there. Go to the section in which you have already found books dealing with the subject and there perhaps other publications will be found which contain the desired data.

Step 7.—Look through trade, financial, and technical magazines. The ones most likely to be productive will be determined by the nature of the subject. Some of these are listed in Appendix A.

Step 8.—If the data are still elusive or perhaps incomplete go through the periodical indexes which are found in the library. The following are most frequently available: *Readers Guide to Periodical Literature*,¹⁰ *Industrial Arts Index*,¹¹ *Public Affairs Information Service*,¹² *New York Times Index*.¹³

⁹ Donald H. Davenport and Frances V. Scott, *An Index to Business Indices*, Chicago: Richard D. Irwin, Inc., 1937.

¹⁰ H. W. Wilson Co., New York.

¹¹ H. W. Wilson Co., New York.

¹² Public Affairs Information Service, New York.

¹³ New York Times Co., New York.

Step 9.—If at this point the desired data have not been found, it is time to consult some experienced person who may have knowledge of them. The experienced person for students means the teacher; for research workers, a fellow-worker or director. Finally, it may be desirable to write to a government or non-government agency for the information. The United States Information Service, 1405 G Street N.W., Washington, D. C., has been established as a Division of the Office of Government Reports to answer inquiries regarding the departments and agencies of the federal government.

Only in the most difficult cases will it be necessary to employ all of these steps. Usually the first two or three will be productive. After a few searches have been made the general contents of the major publications will be sufficiently familiar so that in most cases the proper source can be selected immediately. The further one progresses in the use of library sources the less the need for formal methods and the greater the reliance on experience.

Examples of Library Search

Some problems for library search were assigned to a student who had a slight acquaintance with the titles of the various statistical publications but very little knowledge of their contents. His report of the results of the search is reproduced as Appendix B at the end of this chapter. The report portrays the student's reaction to success or failure during the search with a sincerity which could not have been simulated by the authors if they had attempted to write this appendix.

The examples were arranged so that successive ones would require the use of additional steps of the search process. Careful study of the student's explanations will show that in doing the first few examples he acquired considerable knowledge of the contents of standard sources which saved time in the later examples. This experience and the similar experiences of many other students lead to the conclusion that the only way to acquire familiarity with the contents of published sources is by handling them and searching through them for some definite piece of information.

THE CORRECT USE OF DATA

The search procedure of the preceding section leads to the location of a given set of data in a single source or in two or three alternative

sources. Before the data can be transcribed they must be put through a process of verification and tested for validity.

Verification of Data

The data should be verified (1) to detect discrepancies, (2) by cross-reference when multiple sources are available.

Discrepancies.—Discrepancies in data are usually not difficult to detect but may escape the unwary collector. They may appear as a result of one or more of the following causes.

Changes in unit: Some of the things that may be expected are changes in the unit of measure, changes in the definition of the unit and changes in the nature of the unit. Illustrations of all of these changes can be found in the *Statistical Abstract* for 1936. An example of change in the unit of measure is shown in Table 524 which presents "Imports of Merchandise by Commodity Groups and Articles." On page 536 the first item is wood pulp. The unit used is long tons prior to 1935 and short tons beginning with 1935. A change in the definition of the unit appears in Table 247 entitled "Reporting Member Banks in 101 Leading Cities—Principal Assets and Liabilities." "Demand Deposits Adjusted" is the heading of the next to the last column. Through August, 1934, the data are net demand deposits, but subsequently are adjusted as explained in the footnote. The figures really represent two different things and cannot be regarded as a single series even though they are printed in the same column. Table 426, "Railway Equipment in Service, All Reporting Companies," shows that there was a larger number of steam locomotives in service in 1916 than in 1929 despite the greater volume of traffic hauled during the later year. This is explained by the change in the nature of the thing counted, since a locomotive manufactured in 1916 was hardly the same as a locomotive manufactured in 1929.

Changes in classification: Arrangements according to time, space, or attribute may be involved. A change of the time period for recording railroad data occurred in 1916 when a shift was made from fiscal to calendar years. An adjustment must be made for this change if the earlier and later periods are to be combined in a single series. Changes in the boundaries of the wards of cities have the effect of changing the classification of any data reported by wards. Changes in attribute classifications appear frequently in the biennial *Census of Manufac-*

tures, as illustrated by the following statement introductory to the section entitled "Radio Apparatus and Phonographs."

At censuses taken prior to 1931, the manufacture of phonographs was treated as a separate industry, but the increasing production of radio apparatus by manufacturers of phonographs and the introduction of the combination radio-phonograph unit made it desirable to establish the present classification. Manufacturers of radio apparatus were formerly classified in the "Electrical machinery, apparatus, and supplies" industry. The schedule for this industry did not call for detailed data on the production of radio apparatus, and therefore no comparative statistics are given for years prior to 1931.¹⁴

A discrepancy which is closely allied to a change in spatial classification occurs when the area for which data are reported is changed. Such changes may arise from shifts in national boundaries, in customs districts, or in navigable waters. On the other hand the changes may be of a purely statistical character, as the following examples will show. The Bureau of Labor Statistics report of building permits issued included 262 cities in 1921 and 1922. In subsequent years the number of cities has been gradually increased until in May, 1940, it reached 2,047. The figures are clearly not comparable from 1921 to 1940. Comparable figures over a period of years for 257 identical cities are published in each issue of the *Statistical Abstract*. The birth and death registration area of the United States is another example of changing area. Starting with Massachusetts, New Jersey, and the District of Columbia in 1880, the registration area for deaths has been gradually expanded until in 1933 for the first time all of the states were included. The birth registration area started with ten states and the District of Columbia in 1915 and expanded gradually until all of the states were included in 1933. During this period the number of births and deaths cannot be compared from year to year, but birth rates and death rates are approximately comparable.

Revisions: Perhaps the best example of this type of discrepancy is to be found in *Agricultural Statistics* (formerly included in the *Yearbook of Agriculture*). There are scarcely two yearbooks which give the same series of figures for the country's wheat production. In the issue of 1935 corrections were made as far back as 1866. While there are many other cases of this kind in recorded data, it is unlikely that many can be found which are less stable than the records of crop estimates of the Department of Agriculture. Presumably the only

¹⁴ *Census of Manufactures*, 1933, p. 577.

thing which can be done with such figures is to use the most recent issue and hope that corrections made in subsequent issues will not destroy the validity of the data used.

It cannot be safely assumed, however, that the most recent or revised figure is always correct. Errors in revisions occur less frequently than in preliminary figures, but are more likely to be overlooked. An example appeared in the *Survey of Current Business* during the early months of 1937. The particular series involved was "Total Car Loadings." Table 24 is a reproduction of the data with footnotes intended to explain the changes as printed in three successive issues. The foot-

TABLE 24
TOTAL CAR LOADINGS AS PRINTED IN THE MONTHLY
SURVEY OF CURRENT BUSINESS WITH THE PERTINENT FOOTNOTES
(000 omitted)

AS PRINTED IN THE ISSUE OF	1936		
	JANUARY	FEBRUARY	MARCH
February, 1937*	2,353	3,135	2,419
March, 1937†	2,975‡	3,135	2,419
April, 1937		2,512‡	2,419

*Data for February, 1936, are for 5 weeks, other months, 4 weeks.

†Data for January, 1936, are for 5 weeks, other months, 4 weeks.

‡Revised.

notes do not explain all that happened to this series. In the February issue of 1937, and in the ten preceding issues, car loadings for five weeks were included in the February, 1936, figure, giving 3,135,000 cars. Beginning with the March, 1937, issue the loadings for a week which ended February 1, 1936, were shifted from the February total to the January total. Thus the January, 1936, total was increased to 2,975,000 cars, but an equivalent deduction was not made from the February total. As a result 622,000 cars reported for the week ending February 1, 1936, were counted twice in the March, 1937, issue. The error in the February, 1936, total was corrected in the issue of April, 1937, but unfortunately the March issue is most frequently used because it contains data for the full 12 months of 1936.

Typographical errors: A good example is found in the record of bank clearings printed weekly in the *Commercial and Financial Chronicle*. Individual clearings are printed for more than 100 cities and in that list it is not uncommon to find as many as five changes in the data copied from the previous week. There is no way of knowing

which is the misprint since no explanations are included. Such errors are most likely to occur in publications which are not carefully proofread.

Interruptions in series: Loss of continuity in a series which has been published regularly creates a problem for the user. If the interruption is brief such as the gap in the recording of bank debits caused by the "bank holiday" in March, 1933, simple interpolation may be all that is needed to resolve the difficulty. There are other cases of failure to publish which are less easy to overcome. For example, from July, 1933, to February, 1935, inclusive the Post Office Department found it inconvenient to release for current publication the figures for postal receipts in "Fifty Selected Cities" and in "Fifty Industrial Cities." Such a prolonged suspension brings to a halt any statistical work involving use of the missing data.

Even a slight experience will afford enough background to insure that many of the inconsistencies in published data will be recognized. Beyond that lies the task of detecting the less obvious discrepancies. Two things are necessary for this, the first is varied experience in collection, the second is the exercise of common sense. The latter might be defined as a combination of experience, judgment, and figure perception.

Cross-Reference.—In many cases only one source can be found for a required set of data and no verification by cross-reference is possible. Frequently, however, similar data are collected by several agencies. In these cases all of the sources should be found as a means of determining which is most complete, which contains the data in most usable form, and which has the best general record of reliability.

It is also desirable to get the most recently published source so that any corrections or revisions of the data will be discovered. If the record coincides in all of the sources, that fact gives added confidence in the accuracy of the data. If differences appear, the necessity of reconciling them arises. Discrepancies of the types enumerated in the preceding section may be involved or fundamental differences in the method of collection may be uncovered by study of the notes accompanying the tables. If inconsistencies arise which cannot be explained, it is necessary to search for collateral sources or perhaps to write to the collecting agency for further information.

The process of comparing the data in several sources is known as *cross-reference*. An example of the use of cross-reference will serve

to demonstrate the method and its advantages. Suppose that the following problem were proposed on June 1, 1937: "Collect data on annual production of steel ingots for the years 1932-1936, inclusive." The information obtainable from four sources is shown in Table 25, columns 1 to 4. The four reports contain different figures despite the fact that the original source of all four sets of data is the American Iron and Steel Institute.

The title of the table from the *Statistical Abstract* states that steel ingots and steel for castings are included. Since the problem asks for steel ingots only, these data would not be satisfactory, even though the figure for 1936 could be supplied from current sources. An examination of the March, 1937, *Survey of Current Business* in which the tonnage for steel ingot production and castings is given separately in-

TABLE 25
PRODUCTION OF STEEL INGOTS IN THE UNITED STATES, ANNUALLY 1932-36,
AS REPORTED IN FOUR SOURCES
(thousands of long tons)

YEAR	STATIS- TICAL ABSTRACT* (1)	ANNALIST (2)	STEEL† (3)	STEEL YEARBOOK OF INDUS- TRY** (4)	REVISED SERIES BESSEMER AND OPEN HEARTH PRODUCTION (5)
1932	13,681	13,323†	13,464	13,323	13,323
1933	23,232	22,594†	22,894	22,594	22,594
1934	26,055	25,599‡	25,949	25,599	25,599
1935	34,093	33,426§	33,940	33,418	33,418
1936		46,919	47,513		46,808

*1936 issue, p. 705.

†December 7, 1934, p. 790.

‡February 14, 1936, p. 270.

§April 10, 1936, p. 349.

||February 12, 1937, p. 277.

||May 10, 1937, p. 32, second table, "Annual Steel Ingot Production."

**January, 1937, p. 360, "Steel Ingot Production, 1917-1937."

dicates that none of the other three series, Table 25, columns 2 to 4, includes castings. Further study is needed, however, to reconcile the differences in these three series. The *Annalist* data correspond to the *Steel Yearbook* through 1934, but differ in 1935. If the *Annalist* for October 9, 1936, instead of April 10 is used, revised monthly data are found for all of 1935 which agree with those in the *Yearbook*, leading to the conclusion that the 1936 *Annalist* figures will likewise be revised later in 1937. It can now be concluded that these two series coincide, with only the revised 1936 figure lacking. Headings and footnotes to the respective tables indicate that both include only Bessemer and open-hearth processes.

The data from *Steel*, column 3, are classified in the original source according to processes including crucible and electric as well as Bessemer and open-hearth. The difference between this series and the other two is explained by the inclusion of production by the crucible and electric processes. If a subtotal for Bessemer and open-hearth processes only is computed from the original table, the results coincide through 1935 with those from the *Annalist* and the *Steel Yearbook*. Since the May issue of *Steel* was published later than either the *Yearbook* or the *Annalist*, one can assume its 1936 figure is the more recent revision.

The series for steel ingot production by open-hearth and Bessemer processes can therefore be completed as shown in column 5, and there is now no disagreement among the three sources. There are, however, two complete series to choose from—column 3 which includes crucible and electric production and column 5 which does not. Since the reports issued by the American Iron and Steel Institute usually include open-hearth and Bessemer only, column 5 appears to be the most desirable series to use.

There are two major advantages in conducting this search: (1) the determination of the best figures to use for steel ingot production and (2) the collateral knowledge acquired concerning methods of recording data on steel ingot production.

Evaluation of Data

Evaluation deals not so much with the accuracy of data as with their validity. The question is: Are these data satisfactory for the purpose for which they are to be used? The answer can be obtained by understanding the background of the collection and by visualizing the collection process.

Understanding the Background.—Data come to exist either as a by-product of non-statistical activity or directly for statistical purposes. There are many examples of series of data which are collected for statistical purposes. The work of the Bureau of Census, the Bureau of Labor Statistics, and the Bureau of Agricultural Economics is carried on for the purpose of providing numerical information for general use. The purpose is directly statistical.

Illustrations of data secured as a by-product of other activity are gasoline consumption by motor vehicles and cigarette consumption, both obtained by the Bureau of Internal Revenue in the course of

collecting the taxes levied on these articles by the government. Further examples are a census of employment, which might be tabulated from the registration cards for retirement annuities filed with the Social Security Board by employed workers at the end of 1936, and an index of grocery prices which might be computed from the newspaper advertising of grocery stores.

By-product data are collected for some official or business purpose. Once they have served that purpose the collectors have no further interest or at most only a collateral interest in them. They may be kept in poor form; errors corrected for the major purpose may be omitted from the statistical record; there may be overlaps and omissions which creep in because the statistical record has not been adequately checked; the data may not be in usable form for statistical purposes, although serving the major purpose well. Since the data from by-product sources are likely to contain inaccuracies, it is desirable wherever possible to cross-check them in a direct statistical source.

Visualizing the Collection.—This means asking the question: How were the data collected? By answering this question considerable insight will be gained concerning the difficulties that were encountered in collecting the data and consequently a fair basis may be obtained for judging their reliability. An example will show what is involved in visualizing the collection.

The United States Department of Agriculture publishes estimates of wheat production annually. To collect complete information concerning the amount of wheat produced would involve canvassing each year more than half of nearly 7,000,000 farmers in the United States. This would be a long and costly task and even if it were possible to do the work the results would contain some error because many farmers have no accurate record of the size of their wheat crop. Hence the Department of Agriculture makes no attempt to collect complete data annually. There are crop reporters in all parts of the country who voluntarily send in estimates of the number of acres planted in wheat in the sections their reports cover. Only a small part of the wheat acreage in the country is thus reported, but by applying the estimates to unreported areas statisticians are able to calculate the acreage planted in wheat for the entire country. Then at harvest time the same crop reporters send in estimates of the average yield per acre in their territories. By multiplying the estimated acreage by the estimated yield per acre the approximate production can be obtained for each section

of the country. The total of these sectional estimates is the only annual production figure available for the whole United States. Every five years (ten years, prior to 1920) an actual census of production is taken and the estimates are checked against the census. Table 26 shows that the estimates varied from the census by more than 3 per cent on only two occasions and in four cases have varied by less than 1 per cent. Hence the conclusion is that the Department of Agriculture annual estimates of production are fairly accurate, but the margin of error inherent in the method of collection must be kept in mind when they are used.

TABLE 26

COMPARISON OF DEPARTMENT OF AGRICULTURE ESTIMATES OF
WHEAT PRODUCTION WITH BUREAU OF CENSUS COLLECTION*

YEAR	(1) DEPARTMENT OF AGRICULTURE ESTIMATE (IN BUSHELS)	(2) BUREAU OF CENSUS COLLECTION (IN BUSHELS)	(3) PER CENT VARIATION (1) ÷ (2) - 100%
1879	459,234,000	459,483,000	- .05
1889	504,370,000	468,374,000	+7.69
1899	655,143,000	658,534,000	- .51
1909	683,927,000	683,379,000	+ .08
1919	952,097,000	945,403 000	+ .71
1924	841,617,000	800,877,000	+5.09
1929	823,217,000	800,649,000	+2.82
1934	526,393,000	513,213,000	+2.57

**Agricultural Statistics* (1937), pp. 9-10.

Example of Evaluation.—Table 27 illustrates many of the pitfalls in the use of data and shows the method of evaluating data from the notes which accompany the table.

One quickly detects from reading the several notes that the information contained in this table has variable accuracy. For some states the sales are determined by the number of tags addressed to consumers in that state by fertilizer manufacturers. If the counts are kept accurately, if the bags are all the same size, and if car-load shipments sent to retailers near state boundaries are distributed mainly in the state in which the retailer resides, then the tag count may give fairly good results. For other states estimates are made either by state authorities or by the National Fertilizer Association. Actual records of sales are compiled by state authorities for another group of states. For the year 1929 data collected by the *Census of Agriculture* in 1930 are used as the most reliable estimates of sales in some states but not in others.

TABLE 27

FERTILIZER: ESTIMATED SALES IN THE UNITED STATES

NOTE.—Data are based on fertilizer tag sales for some States and are compiled by State authorities from sales records, etc., for others, as indicated by footnotes. For 1929, census data have been used in many cases. Other figures are estimates made by State authorities or the office of the National Fertilizer Association.

(In tons of 2,000 pounds)

DIVISION AND STATE	1928	1929	1935 (prel.)
United States	7,985,019	8,078,548	6,191,321
New England	365,119	357,465	282,503
Maine	178,750	185,650	125,000
New Hampshire*	16,900	11,500†	16,000
Vermont‡	16,911	14,905	15,295
Massachusetts*‡	70,458	68,611†	63,208
Rhode Island§	10,100	7,909†	12,000
Connecticut	72,000	68,890†	51,000
Middle Atlantic	743,558	798,433	658,874
New York‡	260,000	287,959†	234,000
New Jersey 	143,574	162,361†	149,408
Pennsylvania‡	339,984	348,113†	275,466
East North Central	755,711	820,402	658,696
Ohio‡	320,866	338,662	306,509
Indiana 	221,082	250,201	194,946
Illinois‡	30,509	38,056	23,827
Michigan	150,213	152,812	105,000
Wisconsin‡	33,041	40,671	28,414
etc.			

* Year ended June 30, except data for 1929.

† Agricultural census.

‡ Compiled by state authorities, except as noted.

§ Year ended March 31, except data for 1929.

|| Year ended October 31.

¶ Based on tag sales.

Source: The National Fertilizer Association, *Statistical Abstract* (1936), p. 598.

Certain other peculiarities should also be noted. In New Hampshire, Massachusetts, Rhode Island, and New Jersey there is a discontinuity between 1928 and 1929 data. For example, for New Hampshire the 1928 data cover the period from July 1, 1927, to June 30, 1928, whereas the 1929 data cover the calendar year, January 1 to December 31. Hence the table contains no record of sales in these states during the second half of 1928. A further difficulty appears in footnote ||. Presumably it should read "Year ended October 31, except data for 1929," since footnote † on the 1929 data for New Jersey shows that they are census data and we know that the Census of Agriculture covered the calendar year. Finally, revised 1935 figures are to be found in the 1937 *Statistical Abstract*.

The detailed analysis of the notes accompanying this table indicates the method of evaluating data in terms of the background and sur-

rounding circumstances. Footnotes and headnotes should always be studied carefully to discover what explanations the author of the table believed necessary to its comprehension. To disregard such notes is direct failure to use the available means of verification and evaluation of the data in the table.

Transcribing the Data

The final step in the collection process is to transfer the data from the source to a collection form. Although this appears to be a purely routine matter there are certain rules which, if observed, will help to avoid trouble later. Always assemble all of the data before doing any copying. Too frequently it has been the authors' experience that students bring in part of a series of data and ask for advice on how to complete the series only to find that it cannot be completed and a new series must be found. Until all of the data have been found there is no way of knowing whether partial discoveries will be usable.

In transcribing data always start with the publication of most recent date and work back to the earlier dates. When revisions have been made from time to time, the best way to discover them is by comparing data in the latest publication with overlapping data of an earlier publication. For example, if it is necessary to obtain data from 1929 to 1940, inclusive, and data from 1933 to 1940 are found in one issue of a publication, then the latest issue containing data from 1929 to 1933 should be used to complete the series. Data for 1933 appear in both issues and should be compared to insure that no change has occurred in the recording and that the same series is being taken from both issues.

In another case the data may not agree in the two issues. Then three possibilities arise: (1) Explanations accompanying the tables may state the nature of the revision involved and how to make the series comparable in the two issues. (2) No explanation of the change may be given and it will be necessary to find another source containing the same series in comparable form or a substitute series that will serve the purpose. (3) Failing in both of the preceding expedients, the search may have to be abandoned. Difficulties in matching series in different issues of a source book occur most frequently as the result of shifting the base of an index. Such a revision can usually be adjusted unless an accompanying change in the method of constructing the index entirely destroys the comparability of the two parts of the series.

APPENDIX B

EXAMPLES OF SEARCH FOR DATA IN LIBRARIES

These examples are intended to show how a student ¹⁵ proceeded to find six series of data for problems which were assigned to him.

Example I: Find the monthly freight car loadings by commodity classes for the ten-year period 1927–36.

1. Thought that material would be in the *Statistical Abstract of the United States*, but found that although there were data for freight car loadings, the data were not in monthly form. Made a mental note of this—remembering that most series of figures included in the *Abstract* were indicated by years.

2. Looked in the index to the *Survey of Current Business* and found the necessary material by individual classes and also by months for a particular period. By searching through the back issues or through the *Supplements*, I found the material available for the full ten years.

3. Tried the *Federal Reserve Bulletin* also and found that the same material was included in their monthly issues under the topic of industrial activity.

Example II: Find the monthly indexes of employment and payrolls in the United States for the five years from 1932 to 1936, under this particular heading "Retail Trade—General Merchandising."

1. Discarded the thought of using the *Statistical Abstract*, because I wanted monthly figures.

2. Reached for the *Survey of Current Business* and found under the heading of employment and the heading of payrolls that monthly data were available for "Retail Trade," but that no distinction was made for "Retail Trade—General Merchandising."

3. Because of the nature of the topic, I tried the *Monthly Labor Review* and by chance picked up a monthly issue dated in 1933. Here again the data were available but no distinction was made between General Merchandising and other merchandising. Searched further and found that the January, 1937, issue made this distinction in their current data. In looking through issues for 1936 and 1935, I found the entire series dating back to 1932 in the January, 1935, issue.

Example III: Find the Freight Tonnage Originating on Class I Steam Railways in the United States by quarters from 1927 to 1936. Designate the following commodity groups separately: Products of Agriculture, Animals and Products, Products of Mines, Products of Forests, Manufacturers and Miscellaneous, All L.C.L. Freight.

1. Looked in the *Statistical Abstract of the United States* and found data on freight tonnage, but the data were not in commodity groups nor by quarters. The figures showed tons of revenue freight carried.

2. Tried the *Yearbook of Agriculture* and found data entitled Freight Tonnage Originating on Railways in the United States and also the correct commodity groups, but the data were annual figures.

3. The *Minerals Yearbook* did not have the data in the correct form. And I didn't try the *Survey of Current Business* nor the *Monthly Labor Review*, nor the *Federal Reserve Bulletin* because of their particular use of monthly figures. Of course, this is not always true, but the particular nature of the sources led me to believe that the data were not available in them.

4. Because most of the previous data had been compiled by the Interstate Commerce Commission, I looked in the card catalogue for particular bulletins or statements

¹⁵ Reports written in 1937 by Robert Berner, then a sophomore in the School of Business Administration of the University of Buffalo.

published by the Commission or other independent establishments under the heading of "Commercial and Industrial," or publications of the Department of Commerce. A good index of government publications is the *Monthly Catalogue of Public Documents*. In the 1935 issue, I found under the topic of freight commodity statistics that the Bureau of Statistics of the Interstate Commerce Commission sent out statements quarterly giving Freight Statistics on Class I Steam Railways in the United States which included the total freight tonnage by the commodities indicated.

Example IV: Find the total market value of all listed stocks on the New York Stock Exchange by months from 1926 to 1936.

1. Tried the *Federal Reserve Bulletin* and found certain data about security markets giving the indexes of stock prices by months, but the data required were not found in this source.

2. Did not attempt looking into the *Monthly Labor Review*, *Statistical Abstract*, *Minerals Yearbook* or *Agricultural Yearbook* because of the nature of the subject and problem.

3. Found in the *Survey of Current Business* the stock prices of all stocks, sales, yields, and other information but not the total market values of all listed stocks. At this time, I thought that if I had the base figure and could compute the actual figures for sales and stock prices, that by multiplying the two figures I might have usable figures of market values. This procedure would not be too accurate, however.

4. Looked in the card catalogue for government publications other than the original six, but found nothing dealing with the subject except material which had been used in the *Federal Reserve Bulletin* and the *Survey of Current Business*.

5. Found in the card catalogue that the New York Stock Exchange published a monthly bulletin. Upon getting a copy I found the acceptable material.

Example V: Find the yearly production of steel rails from 1919 to 1936 by the following processes of steel manufacture: open-hearth, Bessemer, crucible, electric, and all others. Compare and explain the variations in each series.

1. Went immediately to the *Statistical Abstract of the United States* expecting to find the yearly figures. I did find figures giving the total rail production from 1914 to 1935, but they did not indicate the four distinct processes.

2. The *Survey of Current Business* contains monthly data on track work production, but does not contain the production according to processes.

3. Exhausted the content indexes of the *Agricultural Yearbook*, *Minerals Yearbook*, *Federal Reserve Bulletin*, and the *Monthly Labor Review*, but found nothing suitable to my needs.

4. Tried the other government sources, looking in the card catalogues and *Index to Government Publications* but found no discrimination in the processes used in producing steel rails. They did contain, at least a few sources contained, the total rail production figures by years, their source being the American Iron and Steel Institute.

5. Picked up a copy of the *Annalist*, finding only the tons of rails ordered by months as taken from the *Railway Age* magazine. This was not satisfying, nor were the data in the *Commercial and Financial Chronicle*.

6. Attempted to find the material in the technical magazine *Steel*. I found the data contained month by month quite inconsistent except for their index of business activity. Each month they include a new set of data recurring irregularly. Looking through several copies, I decided that the material I wanted was not included.

7. Tried the magazines *Railway Age* and *Iron Age* but found nothing satisfactory in either. The data in *Iron Age* are quite consistent, appearing in each monthly issue, showing the last two months, absolute figures. Their index of capital goods is one that is quite widely known and widely used. It shows weekly variations. Many of their figures on steel production and output were taken from the annual statistical report of the American Iron and Steel Institute.

8. I tried this annual report, finding an abundance of data on steel production by processes, one set of which contained the data I wanted.

Example VI: Find the dollar value of department-store sales, annually from 1927 to 1936 for the United States.

1. Glanced through the *Agricultural Yearbook*, *Minerals Yearbook*, and the *Monthly Labor Review*, and as I expected, found nothing pertaining to the required data.

2. Realizing that the *Survey of Current Business* and *Federal Reserve Bulletin* usually contain monthly data, I nevertheless found the data on department-store sales by months in index form. By so doing, I hoped to find the original source of the data. The data, appearing in forms adjusted for seasonal and unadjusted, were compiled by the Board of Governors of the Federal Reserve System, Division of Research and Statistics. These figures represent monthly dollar sales for a sample of approximately 425 stores.

3. Tried the *Statistical Abstract of the United States*, finding the dollar value of department store sales (including mail-order sales) for 1929 and 1933. From the various distinctions given to retail stores, I suspect there is a difficult problem in determining what kind of store may be classified as a department store. The *Abstract* also has an index of yearly department-store sales from 1919 to 1935 which is also a sample of from 400 to 560 stores compiled by the Federal Reserve Board. The actual dollar sales are available for 1929 and 1933 because of the census reports made by the Census Bureau. The original source of these figures comes from the *Fifteenth Census of United States: Distribution*.

4. From the *Census of Business*, I found the dollar value of net sales for 1933 and 1935. This source divided department-store sales into independents, chains, mail-order, commission or company stores, and all others.

5. Found nothing in the *Federal Document Index* nor card catalogue that would lead me to the desired data.

6. The *Industrial Arts Index* in the Buffalo Public Library showed several sources of statistical data pertaining to department-store sales, some of which are included in the foregoing steps. Mr. C. M. Schmalz of the Harvard University Graduate School of Business Administration published a report containing the "Operating Results of Department and Specialty Stores in 1935." Yet these data were not wide enough in scope, either in number of establishments or variety of years.

7. The *Commercial and Financial Chronicle* contains monthly statistics of department-store sales in index form, taken again from the Federal Reserve Board.

8. Looked through various technical sources and card catalogues indicating subjects of technical magazines and books, but found nothing about dollar value of department-store sales, annually from 1927 to 1936.

9. Concluded that data were not available in published sources.

PROBLEMS

1. The answer to each of the following questions is to be found in a commonly used *government* source. (a) Give answers to the questions (as assigned) with exact reference to the source. (b) Describe the steps you followed in each case in order to locate the data.

(1) The percentage of increase in population for the United States and for California, from 1920 to 1930 and from 1930 to 1940.

(2) The total number of strikes in progress in the United States during the month of August of last year; the number of workers involved; and the number of man-days idle during the month.

(3) The wholesale price per bushel of No. 2 hard winter wheat at Kansas City, for the most recent week.

(4) The number of dozen pairs of women's full-fashioned silk hose exported from the United States during June of last year.

2. The answer to each of the following questions is to be found in a commonly used *non-government* source. (a) Give answers to the questions (as assigned) with exact reference to the source. (b) Describe the steps you followed in each case in order to locate the data.
 - (1) The number of new passenger car registrations for Ford, Chevrolet, Plymouth, and Cadillac in November of last year.
 - (2) The number of business failures in retail trade in the United States during the month of May each year since 1939.
 - (3) The percentage of American-made passenger cars sold outside the United States in 1935; motor trucks; total.
 - (4) The gross federal debt as reported by the United States Treasury as of 3 days ago.
3. State which of the steps of library search were employed in each of the examples in Appendix B.
4. Write a report of discrepancies found in the bank clearings reports in certain cross-referencing issues of the *Commercial and Financial Chronicle*. For example, the following issues will serve the purpose: January 18, 1941, February 22, 1941, etc., at monthly intervals.
5. A cursory search and attempted verification of the production of pig iron in the United States in 1937 produces the following figures:

<i>Steel, Yearbook of Industry</i> , January 1, 1940.....	36,709,000 gross tons
<i>Statistical Abstract</i> , 1939	35,224,000 long tons
<i>Survey of Current Business</i> , 1938 Annual Supplement.....	3,051,000 long tons
	(monthly average)
<i>World Almanac</i> , 1939	36,130,000 gross tons
<i>Standard Trade & Securities, Statistical Bulletin</i>	100,300 gross tons
	(daily average)
<i>Penton's Almanack</i> 1940-41.....	41,114,000 net tons

Which of these figures would you choose? Give reasons for your choice by consulting these several sources; explain as completely as possible the apparent discrepancies.
6. From any issue of the *Commercial and Financial Chronicle* select a series of data that are of the by-product variety. Explain the major purpose for which the data were collected. Evaluate the data.
7. Before using library data, what facts would you desire to know about
 - a) the nature of the data themselves? Why?
 - b) the types of units in which the data are expressed? Why?
 - c) the organization collecting or preparing them? Why?
 - d) the purpose for which they were issued? Why?
 - e) the consumers to whom they are addressed? Why?
 - f) the accuracy of the data? Why?
 - g) the homogeneity of the conditions under which the data were collected or to which they refer? Why?

8. Certain difficulties of collection occur in each of the following problems. Find as much information as you can in answering the question and explain the circumstances in the sources that make it difficult to secure complete and comparable data. (The instructor will assign one or more of the problems to each student, according to the time available.)
- a) An important measure of steel ingot production is "per cent of capacity." Trace the changes in "capacity" since 1889.
 - b) Compare the number of savings banks, depositors, and amount of savings in your own state with the United States as of recent date.
 - c) Compare the changes in the number of employees in the carriage industry and in the automobile industry at 10-year intervals beginning with 1900.
 - d) What was the payroll of the executive branch of the United States government annually, 1929 to date?
 - e) Compare the number of full-time employees in one-, two-, and three-store independent groceries in the United States with the number employed in chain grocery stores, in 1929 and in 1935.
 - f) Select the five industries whose indexes of employment were lowest during the most recent month, and compare these indexes with their indexes in 1929 and in 1932.
9. Could Table 21, page 168, and Table 28, page 233, be used for verification by cross reference? Give reason for your answer.

REFERENCES

LANE, MORTIMER B., *How To Use Current Business Statistics*. Washington, D. C.: Government Printing Office, 1928.

The entire pamphlet should be read to obtain background knowledge concerning published statistics. Chapter IV is particularly valuable.

SCHMECKEBIER, LAURENCE F., *Government Publications and Their Use*. Washington, D. C.: The Brookings Institution, 1939.

This book tells how to find government documents. It is useful to statisticians in locating less familiar publications.

CHAPTER XI

RATIOS

THE IMPORTANCE OF RATIOS IN STATISTICS

AMONG all statistical techniques none is so commonly used as the ratio. For instance we speak of having a national debt of \$323 per capita; banks paying 2 per cent interest on savings deposits; a retail merchant making a gross profit of 25 per cent on the cost of goods; sales 10 per cent above those of last year; a death rate of 11.0. Such ratios serve the twofold purpose of (1) simplifying data and (2) increasing their comparability.

Ratios properly presented are so easily understood that an analysis of methods seems almost unnecessary. However, when the student changes from the role of a reader to that of a statistician who must transform primary data into ratio form he finds himself confronted with some problems. There are certain principles that determine the construction, presentation, and interpretation of statistical ratios. An exposition of these will form the content of this chapter and the next.

CONSTRUCTION OF STATISTICAL RATIOS

Statistical ratios are fundamentally the same as the ratios with which everyone becomes familiar in studying arithmetic. In chapter II on "The Use of Numbers," no particular mention was made of ratios, since from the point of view of arithmetic computations they are the same as any other fractions and are handled as such. However, since statistical ratios deal always with concrete values or quantities rather than with abstract numbers, certain modifications of the arithmetic concept of the ratio should be noted at the outset. These include the form of expression, the importance of the item used as the base, the number of units in which the base is expressed and the possibility of relations between unlike as well as like items.

Form of Expression

A ratio in arithmetic is the relation which one number or quantity has to another, its value being expressed as the abstract quotient of the

first divided by the second. The term "ratio" is applied either to the original fraction, to the quotient or to both stated together. In statistics the ratio relationship is always between two precisely defined concrete quantities or values. This relationship is simplified through the division of the first term by the second, but it is never expressed as an abstract quotient. Instead *the value of a statistical ratio is the simplified value of the numerator expressed in relation to one or more units of the denominator.*

For example, the items used in the first ratio quoted were the national debt of the United States of \$42,558,875,571, March 31, 1940, and the population as of April 1, 1940, 131,699,275 persons. Dividing the first number by the second gives a quotient of 323, but this is not the statistical ratio. It must rather be stated as follows: "In 1940 the national debt of the United States was \$323 per capita" or "\$323 for every person in the United States."

The qualifying descriptions of both numerator and denominator items must be either fully stated or clearly understood. When statistical ratios are listed in a table the exact specifications of both numerator and denominator are indicated in the table headings.

Selection of Base

In forming ratios between two abstract numbers either one may be used as the denominator or base, e.g., $5 \div 20 = \frac{1}{4}$ or $20 \div 5 = 4$. Likewise an abstract quotient can be readily understood either in the form .0126, 1.26, or 126; hence there is no occasion for changing the number of units used in the base. However, in every statistical ratio the numbers represent definite concrete items and consequently two questions must always be considered: (1) Which item is the logical base? (2) In what number of units shall it be expressed? These two points must be discussed separately.

The Item.—The denominator of a statistical ratio is always a standard to which the numerator is being compared. The two numbers each refer to concrete values or quantities whose characteristics require that one of them should be used as the standard in terms of which the other is to be measured.

In some types of ratio construction it is immediately obvious which of the two items is the appropriate base:

a) In a comparison between a part and the whole, the whole is always the base.

b) In time comparisons between a recent and a prior recording of like items, the prior event will almost always be taken as the base.

c) In a comparison between an effect and its cause or between two values or events one of which is at least partly dependent upon the other, the cause or the independent item is always the base.

In certain other types of ratios the choice of item for the base depends upon the use that is to be made of the ratio:

a) In comparisons between like totals or between two parts of the same total, either one may be selected as the base according to the emphasis desired.

b) In various accounting ratios such as sales divided by inventory, custom has determined the form that is used.

The Number of Units.—The number of denominator units used as the base may be determined by custom, convenience, or effectiveness. Referring again to some of the first examples quoted in this chapter, the national debt is expressed in terms of *one* denominator unit—so many dollars for each single individual in the population; an interest rate of 2 per cent means two dollars for every *hundred* dollars deposited; the death rate indicates the number of deaths during a given period for every *thousand* persons alive at the beginning of the period.

These examples illustrate the practice of expressing the numerator of a statistical ratio in terms of: (*a*) one unit of the base, (*b*) 100 units of the base, (*c*) other powers of 10 units of the base.

One denominator unit as the base: There are many examples in which the base of a ratio is expressed as a single unit. All per capita ratios use one person as the unit of the base. In agriculture we use production per acre; in railroading revenue per ton-mile and per passenger-mile. The accountant uses a 2 to 1 ratio between current assets and current liabilities as a standard of liquidity, and such examples might be listed indefinitely. The expression of the numerator of a ratio in terms of one unit of the denominator as the base is accomplished by the application of a simple proportion in which x = the desired value for the numerator,

$$\text{numerator: denominator} = x : 1$$

The solution for x requires simply that the numerator be divided by the denominator. The result then becomes a simplified value for the numerator in terms of one denominator unit, similar to that determined for the national debt per capita.

One hundred denominator units as the base: Most of the comparisons made by the lay user of statistics are in terms of per cents. Thus we have a 5 per cent increase in grocery prices, a 3 per cent increase in bank deposits, the grades of a class of students 2 per cent below the average, a selling price which is 130 per cent of the cost, a wheat crop which is 85 per cent as great as last year, humidity of 75 per cent and so on. In each case the number stated as a per cent indicates how many numerator units there are for every hundred denominator units.¹

An illustration of the method of expressing a ratio in terms of 100 units as a base may be taken from Table 21, page 168. Column 1 gives the number of telephones in use during each year from 1931 to 1936. To find the ratio of telephones in use in 1936 to those in 1931 the formula that was used before could be applied but the result would then be $\frac{14,454}{15,390}$ or .94 telephones in 1936 for every one in 1931. Since it is difficult to visualize .94 of a telephone, a base of 100 units instead of one should be chosen.

$$\begin{aligned} 14,454 : 15,390 &= x : 100 \\ x &= 94 \end{aligned}$$

The numerator was divided by the denominator as before but the decimal point was moved two places to the right. The result may be stated: "There were 94 telephones in use in 1936 for every 100 in 1931" or "The number of telephones in 1936 was 94 per cent of the number in use in 1931."

Other powers of ten denominator units as a base: Ten, 1,000, 10,000, 100,000, or even larger numbers of units may be used in the base. An advertiser may state that four out of every ten refrigerators sold last month were "Evercolds," the intention being to express the preference for the Evercold product even more vividly than would be the case if the advertisement stated that 40 per cent were "Evercolds." The use of telephones is expressed in the form, "number of telephones per thousand population." In a published chart dealing with automobile fatalities the following ratios were presented: deaths per 10,000 cars registered, deaths per 100,000 population, deaths per 10,000,000 gallons of gasoline consumed. Fish hatcheries study the propagation of fish in units of 1,000,000 fingerlings planted. The hazards of different

¹ The construction and use of per cents may be reviewed by referring to chapter II.

methods of transportation are expressed by comparing the number of deaths to the number of miles traveled in units of 100,000,000 miles.

Similar usage appears in the field of vital statistics. The death rate for the United States in 1933 was 10.7 per 1,000 persons living in the entire area and the birth rate was 16.5 per 1,000 persons living in the area. When dealing with specific causes of death the base used is 100,000. Thus the 1933 death rate from cancer was 102.2 per 100,000 population in the United States, while the suicide rate was 15.9 per 100,000 population.

There are two rules that determine whether one or some higher power of ten units should be used as the base:

1. The number used as the base should be large enough so that the value of the numerator will appear mainly as a whole number but will have not more than three digits to the left of the decimal point. In Table 28 the figures in column 4 are unwieldy. The rule for significant figures permits carrying these quotients to four or even five digits but one of the advantages of the use of ratios, simplicity, has been lost. The most effective form for these ratios is shown in column 3 in which the results appear as whole numbers of only three digits each.

TABLE 28

ESTIMATED NUMBER OF TELEPHONES IN USE IN THE UNITED STATES, ESTIMATED POPULATION, AND RATIOS OF THE TWO AT FIVE-YEAR INTERVALS, 1920-35*

YEAR	(1) ESTIMATED NUMBER OF TELEPHONES (000 omitted)	(2) ESTIMATED POPULATION (000 omitted)	(3) TELEPHONES PER 1,000 POPULATION	(4) TELEPHONES PER 10,000 POPULATION
1920.....	13,329	106,543	125	1,251
1925.....	16,936	114,867	147	1,474
1930.....	20,201	123,091	164	1,641
1935.....	17,503	127,521	137	1,373

* *Statistical Abstract*, 1936: Telephones, p. 344; Population, p. 10.

2. The number used as the base should be smaller than the number in the original denominator; otherwise the ratio implies more stability than is warranted. That is, a per cent should not be based on fewer used because in each month the base is less than 100. For instance, in than 100 cases. A ratio expressed as so many per thousand should in-

methods of transportation are expressed by comparing the number of

Similar usage appears in the field of vital statistics. The death rate for the United States in 1933 was 10.7 per 1,000 persons living in the

June each failure accounts for 14 per cent; hence one less or one more failure in 1937 would have caused the ratio either to be doubled or reduced to zero.

TABLE 29
NUMBER OF BUSINESS FAILURES IN BUFFALO, NEW YORK,
FIRST SIX MONTHS OF 1936 AND 1937

MONTH	(1) NUMBER OF FAILURES	(2)	(3) PERCENTAGE OF CHANGE IN FAILURES (2)÷(1)—100%
	1936	1937	
January.....	12	2	— 83
February.....	14	5	— 64
March.....	3	7	+133
April.....	8	10	+ 25
May.....	4	11	+175
June.....	7	6	— 14

The same criticism applies to the data in Table 30. A 100 per cent distribution has been computed from 22 cases. The column of per cents immediately conveys the impression that at least 100 accidents were involved whereas it really means that if 100 accidents had occurred about 36 of them would have been caused by mine cars, the estimate being accurate only to the extent that a prediction can be based on an experience of 8 cases out of 22. The computation of per cents to two decimal places in this table is further cause for criticism. It is spurious accuracy because the transfer of one accident to a different class (the minimum change possible in the table) would result in a change of 4.5 points in the affected ratios. For example, if there had been 9 fatal accidents due to mine cars and 9 in the miscellaneous class, then the per cent in each of these two classes would be changed to 40.91 per cent. Obviously there is no reason for carrying per cents to even one decimal place when they are based on so few cases.

TABLE 30
FATAL ACCIDENTS AMONG OUTSIDE WORKERS AT BITUMINOUS COAL MINES IN
PENNSYLVANIA, CLASSIFIED BY CAUSE, 1924*

CAUSE OF ACCIDENTS	NUMBER OF ACCIDENTS	PERCENTAGE DISTRIBUTION OF ACCIDENTS
Mine cars	8	36.36
Railroad cars	1	4.55
Electricity	3	13.64
Miscellaneous	10	45.45
Total	22	100.00

* *Pennsylvania Departmental Statistics* (Commonwealth of Pennsylvania, Department of State and Finance, Harrisburg, Pennsylvania, 1925), p. 139.

Kinds of Ratios

The basic definition of an *arithmetic* ratio includes the qualification that the two quantities must be of the same kind and expressed in the same unit of measure. It follows that whenever data are homogeneous they provide suitable material for ratio comparisons. However, before proceeding with the discussion of ratios between items of the same kind, an explanation is necessary concerning *statistical* ratios that are made up of unlike items.

Ratios between Unlike Items.—The possibility of such ratios in statistics was indicated in the first section of chapter VIII in the statement that one type of table may contain "several sets of information . . . not expressed in the same unit, but they . . . *bear some relation to each other.*"

Arithmetic textbooks say, "We cannot express the ratio of a horse to a sheep," and "No ratio exists between five tons and 30 days." Yet even a brief experience in statistics shows that it is exactly such pairs of unlike items usually expressed in two different units that do provide the material for many statistical ratios in common use. Examples are: the rate of production per day or per acre, the income per capita, freight revenue per mile of railway, or bad debt losses per dollar of sales.

Such ratios are permissible in statistics because, as previously noted, the statistical ratio is not an abstract quotient. Dollars of revenue are not actually divided by miles, nor bushels by acres. The statistical ratio is merely a simplified statement of a factual relationship that does exist in each case between numerator and denominator items. For example, the total number of bushels of wheat that is produced depends upon the total number of acres under production; hence it is justifiable to divide the first number by the second in order to arrive at a simpler figure which will indicate the average number of bushels produced per acre cultivated.

Careful scrutiny is necessary in many cases in order to ascertain whether the items that are being compared are really like or unlike. This is particularly true of items measured in *dollar* value. The dollar appears to be the same unit whether it represents dollars of credit or dollars of sales; hence dollar values are readily combined in ratios. In other instances, a word such as "persons," "products," etc., may be used in both terms of a ratio, but unless the word is defined identically in the two terms the ratio is between unlike items and is subject to definite limitations.

In the construction and use of ratios between unlike items whether expressed in the same or different units there are three points of caution to be observed.

1. *The numerator and denominator items although representing different objects or values must be identically defined in both time and space.*

Table 21, page 168, contains three different sets of data: telephones, messages, and revenue, each of the latter two being subdivided according to attribute. All three have the same time classification in the stub and since there is no space classification the spatial characteristic is identical for every item in the table, that is, each represents territory served by the Bell Telephone System. In every horizontal row, therefore, the time and space characteristics are identical for all three sets of items and ratio relationships may be looked for between them. Thus in 1931 the ratio of the number of local messages to the number of telephones was $\frac{22,705,000}{15,390}$ or 1,475 messages per telephone; the total revenue per telephone in the same year was $\frac{\$1,050,200}{15,390}$ or \$68; the average payment per toll call in the same year was $\frac{\$326.30}{985.5}$ or \$.33.

Obviously one would not compare the number of telephones in 1931 to the revenue for a different year nor if state data were available would the number of messages in New York State be compared to the total number of telephones in New Jersey. A complete table such as this one is not always available when single ratios are being used, but it is always possible to reconstruct the table headings in outline in order to test whether the unlike items being used in any given ratio relationships do conform to this rule of identical time and space for numerator and denominator.

2. *There must be a very definite relationship, causal or otherwise, between numerator and denominator.*

In each of the three ratios between unlike items quoted in the preceding paragraph the numerator is in some degree dependent upon the denominator item. The messages are dependent upon the telephones because the telephones must be used in transmitting the messages; operating revenue comes into existence only if telephone instruments are in use; and the revenue from toll calls arises from the fact that the toll calls have been made.

It is easy to assume ratio relationships in such cases as these without

giving enough care to the definitions of the items used. The use of general terms may be correct in some ratios, whereas in other cases that appear to be similar a more specific term is needed to bring out the desired relationship. For example, the ratio "population per automobile registered" might be used in measuring traffic density, but if the standard of living is being measured the ratio should be "registered *passenger* automobiles to population."² These two ratios also illustrate the point that the purpose for which a ratio is being used will determine which of the two items is dependent upon the other. The fact that a certain item such as population is used as the base in the majority of the ratios in which it occurs does not prove that it will be the base item in every case.

3. *The relation between numerator and denominator must be correctly expressed.*

A full and accurately worded statement of the relationship is important in any type of ratio but especially so when the numerator and denominator represent different things. In particular, one is never a "per cent of" the other since in the case of unlike items one could not be any number of 100ths of the other. If the base is conveniently expressed in 100 units, the use of "per cent" is permissible, provided it is combined with "number," "value," or some corresponding expression. For example, "The *number* of teachers is 20 per cent of the *number* of students," or "There are 20 per cent *as many* teachers as students," but certainly not "The teachers are 20 per cent of the students." An experiment to determine the effect of fertilizer upon wheat yield showed that the yield increased four bushels per acre when 100 bushels of lime were spread per acre. This statement involves two ratios between unlike items. Clearly it would be incorrect to say there was a 4 per cent increase either of the lime or of the wheat.

Ratios between Like Items.—Statistical data are considered "like" if they are expressed in the same unit and differ with respect to only one characteristic, according to the classifications that were used in the original tabulation. They may be alike in all attributes and in time, differing only in space; they may be identical in attributes and in space but different in time; or they may be alike in both time and space and in all but one attribute. This last group can be distinguished from the "unlike" data discussed in the preceding section which are also identical

² See chapter XII for further discussion regarding refinement of definition in the construction of such ratios.

in time and space. Unlike items may even be expressed in the same unit of measure but they are differentiated from one another by separate definitions, which show that the items are in different categories.

Referring again to the table of the Bell Telephone System, columns 4, 5, and 6 may be considered separately as a two-way table of like items. Total operating income in the entire system is subclassified according to attribute, local and toll, and is cross-classified according to time. Thus *in any one row*, the data in columns 4, 5, and 6 are alike in time and space and differ only in the one attribute according to which operating income has been subdivided. They are, therefore, "like" and may be compared. We may say, for example, that in 1931 the revenue from local service was $\frac{723.9}{1,050.2}$ or 68.9 per cent of the total operating income. Similarly any figure in columns 4, 5, or 6 may be compared with another figure *in the same column*. They are alike in space and in attribute, differing only in time. Thus the toll revenue decreased from \$326,300,000 in 1931 to \$243,900,000 in 1933, a decrease of \$82,400,000 or 25.3 per cent.

Items that are listed under a single heading in a table are often potentially subject to further subdivision. For example, a single set of data headed "United States" might be subdivided according to the main geographic divisions, according to Federal Reserve Districts or according to the 48 states. "Total wage earners" might be subdivided into male and female. They might also be subdivided into age groups or by wage rates. The danger of comparisons between items that are too general in definition has already been noted in the case of ratios between unlike items, and the warning is equally applicable to ratios between like items. When subclassifications or refinements in definition are available, the maker of ratios should proceed with care before he looks for relations between general data that appear to be "like." However, his refinements can go no farther than the available data will permit. If, according to the classification used, the items are like in every characteristic but one, then they may be combined in ratios. But the possibility that the relations between such data might be affected by further subdivision of their characteristics must be kept constantly in mind in drawing conclusions from these ratios. This point becomes of special importance when comparisons are drawn between two or more ratios and will be discussed further in a later section.

In conformity with the fundamental classifications of data, a ratio between like items may be classified as a time, space, or attribute ratio according to the one respect in which the numerator item differs from the denominator. A second method of classifying a ratio is according to whether (1) the numerator item is a part of the denominator item; (2) the numerator and denominator are separate parts of the same total; (3) the numerator and denominator items are separate totals. The mechanics of ratio construction will be discussed according to these part-total relationships and in each case any differences in the treatment of time, space, and attribute ratios will be pointed out.

Part-to-total: This type of ratio is used chiefly in space and attribute comparisons. Items that differ in time also become material for part-to-total ratios when they are of such a nature that they make up a cumulative total, as for example monthly production figures for a given year. The method of construction and use of part-to-total ratios is identical for all three types of comparison.

Part-to-total ratios take two common forms: (1) the comparison of a single part to the whole and (2) a percentage distribution in which all the parts are shown as percentages of the whole or 100 per cent.

Single Ratios: Examples of single part-to-total ratios are the percentage of manufactured products in the state of Michigan produced in the Detroit area; the number of factory workers over 65 years of age per 1,000 factory workers and the number of high-school graduates entering college per 1,000 high-school graduates. In each of these examples the part selected for comparison with the total is chosen to demonstrate a particular point and so far as that demonstration is concerned nothing need be known about the other parts of the total except that they exist. The first example was a spatial ratio and the others were attribute ratios.

When part-to-total time ratios are to be constructed, a distinction must be made between series in which there is no overlapping between the separate items and those in which the quantities or values do overlap. In the first type the separate parts can be added to make up a total for a longer period of time, as for example, the sum of the exports for each of 12 months in a year will equal the total year's exports and it follows that the data may be used in part-to-total ratios. Series like this are quite different in nature from those which are recorded at similar time intervals but which represent overlapping quantities or values. Such time series as number of employees, number of acres

under production, population or assessed value of property cannot be added to form totals. Consequently no part-to-total ratios can be constructed from them.

The telephone table (Table 21, page 168) may again be used to illustrate the contrast in these two kinds of time series. Columns 2 and 3 show the number of messages of a certain kind that were transmitted during each year from 1931 to 1936. Here there is no overlapping—every message counted in 1931 is distinct from those counted in each of the other years. If there were any special significance in the six-year period, the number of messages of each kind could be totaled and the ratio of any one year to the total period could be used. Columns 4, 5, and 6 which show the operating income for each year likewise consist of non-overlapping items and could be treated in the same way. However, the items in column 1, the number of telephones in use during each year, cannot be added to give a total. They are obviously overlapping data, since most of the 15,390,000 instruments in use in 1931 are also counted among the 13,793,000 used in 1932. Some new ones have been added while some of the old ones have been disconnected and like changes have occurred every year. Since the separate items do not constitute a total, no part-to-total ratios can be made from them. The only possible ratios would be those between two single figures in the same column, that is, total-to-total ratios. Time ratios of this kind will be discussed in a later section.

Percentage Distributions: The same types of data that are suitable for single part-to-total comparisons can be presented as percentage distributions. This is a ratio technique that gives emphasis to the relative importance of each of the parts that make up a total. The several numerator items are each expressed in terms of 100 units of the same denominator, the denominator being equal to the sum of the numerator items. Table 31 shows the amounts loaned on non-farm mortgages by different types of lending institutions during five months of 1939 with a percentage distribution of the several items. The per cent column shows more clearly than the original data that savings-and-loan associations were the most important lending agencies during this period, and that banks and trust companies were second. The least business was done by insurance companies and mutual savings banks with 8.8 per cent and 3.2 per cent, respectively, or only 12 per cent for the two combined.

Table 32 presents a part-to-whole analysis which resembles the pre-

TABLE 31

NON-FARM MORTGAGE RECORDINGS IN THE UNITED STATES BY TYPE OF
MORTGAGEE, FIRST FIVE MONTHS OF 1939*

TYPE OF LENDER	VALUE OF MORTGAGE RECORDINGS (000,000 omitted)	PER CENT OF TOTAL RECORDINGS
Savings-and-loan associations	\$431.8	30.1
Insurance companies	127.1	8.8
Banks and trust companies	359.2	25.0
Mutual savings banks	46.7	3.2
Individuals	263.7	18.4
Others	208.8	14.5
Total	\$1,437.3	100.0

* *Federal Home Loan Bank Review*, Vol. 5, No. 10 (July, 1939), p. 311. Federal Home Loan Bank Board, Washington, D. C.

ceding one in form. At first glance it might appear to be another percentage distribution—in this case of items differing in space. The percentage living on farms varies from 2.4 in Rhode Island to 31.4 in Vermont and a total of these per cents happens to be about 100, which might be assumed to represent the total for the New England and Middle Atlantic States. However, a closer inspection shows that this is not a percentage distribution but a series made up of the first type of part-to-whole ratios. The separate ratios are not comparisons in space but of attribute, i.e., residence on farms is an attribute of a part of the population of each state. Each of the ratios has been computed from a different base, the total population of that state; therefore they cannot be added to give a total that has any meaning. The percentage of the total population living on farms in all of the states together must be computed from the total original data, the same as was done for each separate state.

TABLE 32

PER CENT OF TOTAL POPULATION LIVING ON FARMS IN NORTHEASTERN STATES,
1930 CENSUS*

STATE	PER CENT LIVING ON FARMS
Maine	21.4
New Hampshire	13.5
Vermont	31.4
Massachusetts	2.9
Rhode Island	2.4
Connecticut	5.4
New York	5.7
New Jersey	3.2
Pennsylvania	8.9

* *Statistical Abstract*, 1936, p. 8.

Errors To Avoid in Percentage Distributions: Sometimes per cents of a total are quoted that do not amount to 100 per cent, usually due to some kind of carelessness. This error can be avoided if all the per cents are quoted in tabular form, including the 100 per cent total.³

Other examples, such as Table 33, may be found in which the total of a percentage distribution greatly exceeds 100, not because of error in the computations but because the table contains two distributions instead of one. In this case further confusion is added because in each of the two the data have been distributed in a double classification. Thus there is no clear distribution according to each characteristic separately. In the source from which Table 33 was taken the percentages and primary data were all given in a single column which was even more confusing than in the form shown, since there was no indication which items together totaled 100 per cent. A more usable form for the data is shown in Table 34 which is practically equivalent to two separate tables. In this form comparisons are immediately apparent between the percentages normal and defective in the various categories.

TABLE 33
CLASSIFICATION OF DEFECTS BY SEX AND NATIVITY
FOURTH-CLASS SCHOOL DISTRICTS, PENNSYLVANIA, 1917-18*

	NUMBER	PER CENT
Total male	240,553	
Normal	55,735	11.5
Defective	184,818	38.1
Total female	244,455	
Normal	63,858	13.2
Defective	180,597	37.2
Total native	464,034	
Normal	115,671	23.9
Defective	348,363	71.8
Total foreign	20,974	
Normal	3,922	0.8
Defective	17,052	3.5

* *Departmental Statistics* (Commonwealth of Pennsylvania, Department of State and Finance, Harrisburg, Pa., 1925), p. 72.

This same type of error may appear in a number of different forms, in all of which the mistake lies in the attempt to show too much in one distribution. Percentages of subtotals should not appear in the same column with percentages of the total distribution unless italicized or otherwise unmistakably distinguished. It is preferable to make

³ For a discussion of significant figures in percentage distributions refer to chapter VIII, pages 168-69.

several short tables, each showing one set of relationships clearly. Other isolated percentage relationships that do not warrant the construction of a special table may be pointed out in the text and in any such case the original data should be quoted along with the per cent.

TABLE 34

PUPILS IN FOURTH CLASS SCHOOL DISTRICTS IN PENNSYLVANIA: NUMBER AND PER CENT NORMAL AND DEFECTIVE ACCORDING TO SEX AND NATIVITY, 1917-18

	SEX			NATIVITY		
	Male	Female	Total	Native	Foreign-Born	Total
NUMBER						
Normal.....	55,735	63,858	119,593	115,671	3,922	119,593
Defective.....	184,818	180,597	365,415	348,363	17,052	365,415
Total.....	240,553	244,455	485,008	464,034	20,974	485,008
PER CENT						
Normal.....	23.2	26.1	24.7	24.9	18.7	24.7
Defective.....	76.8	73.9	75.3	75.1	81.3	75.3
Total.....	100.0	100.0	100.0	100.0	100.0	100.0

One of the most frequent misuses of a percentage distribution results from the inclusion of a miscellaneous class. Such a class may contain (1) items which are known to be distinct from those included in the separate classes or (2) items that are unknown or poorly defined.

1. If a class is designated as "All other," "Others," or "Not elsewhere classified," it indicates that a number of less important classes of the distribution have been combined in order to conserve space, to concentrate the reader's attention on the important items or to avoid disclosing confidential information. The characteristics of all these other items are known and they definitely do not belong in any of the specifically named classes. No single class included among "Others" should be larger than the smallest class that is named separately, although the total of the combined "Others" may be greater. In Table 31, "Others" presumably includes endowment funds, non-profit institutions, etc., each of which is distinct from and less important as a mortgage investor than the separately listed lenders of the table. Under such circumstances the information contained in the specific classes loses none of its accuracy by reason of the inclusion of a

miscellaneous class. A percentage distribution that includes the "Others" as one of the parts of the 100 per cent total will therefore correctly represent the relation of each part to the total.

2. In tabulating primary data it frequently happens that the answers to certain questions are missing from some of the collected schedules. Faulty questionnaire planning may likewise result in a group of poorly defined answers that cannot be classified precisely. Such cases must be grouped in an "Unknown" or "Not reported" class, although at least some of them should have been included in one or more of the known classes. The calculation of a percentage distribution with this unknown group as a component part of the total would therefore distort the true relation of each of the specific groups to the total. An alternative method of dealing with this situation will depend upon the circumstances surrounding the collection of the data.

a) If it can be assumed that the known cases comprise a representative sample of the total, the unknown group even if relatively large may be dropped and a percentage distribution computed of the total known cases. This is justifiable in any case if the unknown group is relatively small, since the omission of a few items from one or more groups will not materially affect the percentage relationships. A footnote may be added stating the number of items omitted and what per cent they are of the total number investigated.

b) If a large unknown group has resulted from some element of bias in answering the questions, the distribution of known items can not be assumed to be representative. In such cases no percentage distribution should be computed and indeed the original data are of questionable value.

Table 30 illustrates a so-called miscellaneous class that is really unknown. The source from which the table was taken gave no direct or collateral information to indicate whether the ten accidents classified as miscellaneous were attributable to causes other than those listed or whether several of them were not allocated because of insufficient information. If the cases in the miscellaneous class are independent of the listed causes, then none of them should be more important than the listed causes. Since the table lists one accident involving railroad cars it would follow that ten different causes of one accident each are included in the miscellaneous class. While this situation is quite possible it seems more likely that these ten accidents have been grouped in a miscellaneous class because of insufficient information to allocate

them. If this is the correct interpretation, then the entire table is worthless because the allocation of these ten cases to specific causes might change completely the distribution of cases in the three classes.

Part-to-part and total-to-total: Ratios between two like items neither one of which is a total including the other may be either part-to-part or total-to-total. From the point of view of ratio construction they may be considered together since there is no essential difference in method.

In the case of space ratios the difference is only in the point of view. The areas of Canada and the United States may be regarded as separate totals or they may be two of the component parts of the total area of North America. From either viewpoint the area of Canada to the United States is in the ratio of 106 : 100, or it is 106 per cent as great as the area of the United States.

In time series when the data are non-overlapping they may be regarded either as separate totals or as parts of a larger total; if they do overlap they are always separate totals. However, the method of comparing one item with another or with an average or other standard is the same in either case.

Attribute ratios may appear to be comparisons between separate totals but if they are made up of genuinely "like" items a broader definition can be found under which they will range themselves as two component parts of a larger total. If the two items that are being compared can in no sense be regarded as mutually exclusive parts of a total, then they are not attribute ratios of like items, even though they appear to be expressed in the same unit. They are instead ratios between unlike items and are subject to the limitations already mentioned under that head. For example, the results of a study of radio advertising yielded the following sets of data: total number of persons interviewed; number who listened to a given radio program; number who bought the product advertised on the program. All three of these sets of data used the general unit "persons." This unit had been subdivided in two ways: listeners and non-listeners; buyers and non-buyers. Relationships between listeners and non-listeners, buyers and non-buyers or listener-buyers and listener-non-buyers were genuine part-to-part ratios. But "total listeners" and "total buyers" were not mutually exclusive categories under the general term "persons." Hence for the purpose at hand they were unlike items. A ratio between them would have been valid only if the number of one group were in some way dependent

upon the number in the other group, an assumption that would have been difficult to justify.

Usefulness of Part-to-Part Ratios: Ratios between the several parts will frequently provide more exact information than the ratios of each part to the total. In the field of vital statistics such ratios as the number of male births to the number of female births, foreign-born to native, urban to rural, and white to colored population are in common use. In these cases the corresponding part appears to be a more natural standard than the total of the two. Furthermore, the use of a small base emphasizes the degree of difference between the two parts more effectively than if each were compared with the total.

The part-to-part ratio is equally advantageous in the field of business. Table 31 showed that out of every \$100 of new mortgage loans \$30 were made by savings-and-loan associations, and \$25 by banks and trust companies. A part-to-part ratio would afford the more direct statement that only \$83 was loaned by banks and trust companies for every \$100 by savings-and-loan associations. Or a statistician employed in a mutual savings bank might state that for every \$100 loaned by that type of bank \$925 was put out by savings-and-loan associations.

This example brings out the point that the purpose of such ratios is the comparison of one item to another as a standard of measure; therefore either item may be used as the base according to the emphasis desired. Whether the part-to-part relation has greater significance than part-to-total will also depend upon the emphasis needed in each case. This becomes especially important when two or more sets of such ratios are being compared, usually at different periods of time.

Percentage Relation: Since part-to-part ratios as well as part-to-total ratios are usually expressed in terms of per cents, precise terms must be used in expressing either kind of ratio in order to avoid ambiguity or misstatement. Furthermore in stating a part-to-part relationship, one item is no more a "per cent of" the other than is the case with ratios between unlike items. The sales of chain grocery stores and of independent grocery stores in a community for a given year might appear as follows:

Chain-store sales	\$250,000
Independent-store sales	200,000

If a statement were made, "The independent-store sales were 80 per cent," this could be taken to mean 80 per cent of the two combined. "The independent-store sales were 80 per cent of the chain-store sales" would imply that independents were a part of the chains instead of an entirely different type of grocery. "The relation between the two is 80 per cent" fails to indicate which one is used as the standard. The following are some of the correct statements that can be made: "Independent-store sales were 80 per cent as great as chain-store sales;" "Sales in the chain stores amounted to 125 per cent as much as the independent-store sales."

Percentage Difference: The relation between two parts is very frequently expressed as a percentage difference and may be computed by either of two methods: (1) by subtracting 100 per cent from the percentage relation, computed on either item as base or (2) subtracting the item selected as base from the numerator and dividing the remainder by the base item. Due regard must be taken throughout for algebraic signs according to either method. Using the same example of grocery store sales:

$$\begin{array}{l} 80 \text{ per cent} - 100 \text{ per cent} = -20 \text{ per cent} \\ \text{or} \quad 200 - 250 = -50 \text{ and } \frac{-50}{250} = -.20 \text{ or } -20 \text{ per cent.} \end{array}$$

Again the wording must be precise and the base must be clearly indicated. "The difference between chain-store sales and independent sales was 20 per cent" does not tell which type of store has been used as the base or which had the greater sales. "Sales in independent stores were 20 per cent less than in chain stores" is a much clearer statement; or, if independent stores are selected as the base, "Sales in chain stores were 25 per cent greater than in independents," or "exceeded sales in independents by 25 per cent." Note that whenever the base is changed the percentage difference will change in amount as well as in direction.

Precision of statement is particularly necessary when the part-to-part or total-to-total ratios are time relationships. Differences between two items that are identical except in time are best expressed as per cents of *positive* or *negative change*, or per cents of *increase* or *decrease*, the methods of computation being the same as for deriving percentage *difference* in space and attribute ratios.

Table 35 provides examples of a number of time ratios in each of which an item in October, 1937, is compared with an identically

defined item in October, 1936. All of these examples are total-to-total comparisons rather than part-to-part, since the two months compared are corresponding parts from different years instead of parts of the same year. The first four indicators are non-overlapping series but the fifth represents overlapping data. Despite this difference the same kind of wording can be used in reading all of the ratios in column 4: "In October, 1937, the production of steel ingots showed a 25.2 per cent decrease in comparison with the same month of the previous year"; "The number of cotton spindles active in October, 1937, showed but little change since October a year ago, an increase of only .3 per cent."

TABLE 35
INDICATORS OF BUSINESS ACTIVITY, OCTOBER, 1937 AND OCTOBER, 1936*

BUSINESS INDICATOR	(1)	(2)	(3)	(4)
	AMOUNT OR VALUE		PERCENTAGE RELATION (2) ÷ (1) × 100	PERCENTAGE OF CHANGE (2) — (1) ÷ (1) × 100 OR (3) — 100%
	Oct., 1936	Oct., 1937		
Steel ingot production (thous. tons)	4,534	3,393	74.8	— 25.2
Domestic auto. sales (Gen. Mot.) (number)	44,274	107,216	242.2	+ 142.2
Bituminous coal production (thous. tons)	43,321	40,040	92.4	— 7.6
Building contracts (mill. dollars)	226	202	89.4	— 10.6
Cotton spindles active (thous. spindles)	23,662	23,724	100.3	+ .3

* *The Annalist*, Vol. 50, No. 1295 (November 12, 1937), pp. 796-97 and Vol. 50, No. 1296 (November 19, 1937), pp. 836-37. New York Times Co., New York.

If in Table 35 the comparisons had been with the previous month the first four might be regarded as part-to-part ratios but the wording would be no different except that September, 1937, would be named instead of October of the preceding year. Frequently the period used as a standard in constructing such ratios is not clearly indicated. A newspaper headline may read, "Department-Store Sales Jump Seven Per Cent in August," but careful reading of the article discloses that the 7 per cent gain was not since July of the same year, as one might assume, but since August of the preceding year. In this connection it should be noted that in comparing two time ratios, both of which are based on the same previous standard as 100 per cent, these ratios or "index numbers" are handled exactly as if they were primary data. That is, the relation or difference is found by dividing one by the

other, *not* by subtracting one index from the other. In the example mentioned the index of department-store sales rose from 83 in August, 1938, to 89 in August, 1939, an increase of 6 points on a base of 83 which is at the rate of 7 per 100 or 7 per cent.

Distinguishing between Percentage Relation and Percentage Difference: There is seldom any difficulty in distinguishing between percentage relation and percentage difference in either space, attribute, or time ratios provided the difference is less than 100 per cent. When the difference is large, it is easy to forget that one item must be subtracted from the other to obtain the percentage difference. Using the previous example of grocery-store sales, suppose that two years earlier the sales of the chain stores were only \$25,000. Then the sales for the later year divided by the sales for the earlier year equal + 10 or + 1000 per cent. The sales in the later year therefore were 1000 per cent *as great as* the sales in the earlier year. To obtain the *difference*, 100 per cent must be subtracted, leaving an *increase* of 900 per cent. Or, find the difference between the two years, that is, the later or base year minus the earlier year and divide by the earlier year: $\$250,000 - \$25,000 = +\$225,000 \div +\$25,000 = +9$ or 900 per cent increase. Further illustrations can be found by comparing column 3 with column 4 in Table 35.

As has already been indicated the base item in time ratios is practically always the earlier period. Failure to observe this rule leads to still further confusion in the expression of percentage relation or per cent of increase or decrease, as illustrated in the following quotation:

Making Hilarity Pay.—The large majority of the bootleggers have now cut their prices from 200 to 300 per cent in a desperate effort to meet the competition of the State Liquor Stores.—(Newspaper clipping.)

The reader would assume from the word "cut" that an earlier period had been used as the base. Whatever the former price may have been, a cut of 100 per cent would reduce it to zero, hence any greater decline would mean that the bootleggers were paying the purchasers to take their wares. A decline or decrease can never exceed 100 per cent. Very likely what happened was that liquor formerly selling at \$3.00 per quart was reduced to \$1.00 or \$.75. The difference of 200 or 300 per cent was found by using the later period as the base of the ratio. Assuming that the present price is \$.75, the method should have been as follows: $\$.75 \div \$3.00 = .25$ or 25 per cent. Subtracting 100 per

cent leaves — 75 per cent. Thus the present price is $\frac{1}{4}$ the former price or has decreased 75 per cent. There were two errors in the quoted statement: (1) the later instead of the earlier year was used as the base in computing percentage change and (2) the difference was incorrectly interpreted as percentage decrease instead of the per cent by which the past exceeded the present. There are a few occasions on which a later period may be used as the base, as when we say that the output of a plant was 10 per cent higher last year than this, or pre-war prices were 20 per cent below the current level, but examples of this kind occur so infrequently that they probably would better be disregarded entirely because they tend to confuse the unwary.

PRESENTATION OF RATIOS

Considerable attention has been devoted to proper presentation, both in text and tabular form, during the discussion of the construction of various kinds of ratios. These rules need only be reviewed briefly together with some reference to chapter VIII and with certain additions regarding ratio presentation in general.

In Text

The following points should be observed in any textual reference to ratios:

1. The exact scope of both numerator and denominator should be fully defined unless very clearly understood.
2. The expression of each ratio should be precisely and accurately worded according to the kind of relationship involved, leaving no possibility of misunderstanding.
3. If a ratio that does not actually appear in an accompanying table is used in the text, the data from which the ratio is derived should be quoted along with it.

In Tabular Form

The following rules will be a guide in tabular presentation:

1. The rule of definite and adequate headings in presenting primary data in tables also applies to ratios. If the original data are not included in the table the numerator and denominator items as well as the direction of relationship between them must be clearly defined in

the table. If the data are listed in parallel columns, they may be referred to by column number, as in Table 35.

2. A separate derivative table should be made for each type of ratio comparison drawn from a given set of primary data. In particular, percentage distributions according to both horizontal and vertical classifications or according to more than one category should not be presented in a single table.

3. Every percentage distribution should include a 100 per cent total and the separate per cents must add to 100 except for having been rounded off according to the rule for significant figures. Carrying per cents to too many decimal places gives a false impression of accuracy.

4. Percentages of difference or change must be clearly indicated as positive or negative.

5. Whenever possible, the data from which ratios have been derived should be shown along with the ratios.

Importance of Including Original Data

The last-named point is of importance in presenting any type of ratio. In a complete presentation of any subject the original data will, of course, appear in primary tables and need not necessarily be repeated in every derivative table. In a small summary table, however, there is little danger of too great complication if the data and ratios are arranged in parallel columns. For example, in Table 32 the meaning of the percentages would have been much more evident had two additional columns been given as follows: "Number of Families in State" and "Number of Families Living on Farms."

It should be remembered that the reader is rightly skeptical in accepting any statement of relationships that he cannot verify by making the computation himself. Table 36 contains a number of errors, but because the original data are given along with the ratios, it is possible for the reader to detect the errors and to correct them as well as to add his own interpretation.

Two of the errors in Table 36 are typographical, such as may often be found as a result of lack of careful proofreading. One of these must be discovered in order to avoid misinterpreting the percentage distributions (note also the incorrect caption of this column), whereas the other is less serious. The first line of the distribution for June 30, 1933, as printed is really the first line of the distribution for June 30,

TABLE 36

AMOUNT OF PUBLIC DEBT DUE BEFORE AND AFTER JAN. 1, 1939, EXCLUDING
PRE-WAR, POSTAL SAVINGS AND UNITED STATES SAVINGS BONDS AND
SECURITIES ISSUED EXCLUSIVELY TO GOVERNMENT AGENCIES AND
TRUST FUNDS

DIVISION	AMOUNT (IN MILLIONS)	P.C. OF Tot.
JUNE 30, 1932		
Due before Jan. 1, 1939.....	\$10,870.7	60.2
First Liberty bonds (1947), called 1935.....	1,933.2	10.7
Due after Jan. 1, 1939.....	5,258.8	29.1
Total	\$18,062.7	100.0
JUNE 30, 1933		
Due before Jan. 1, 1939.....	\$13,458.4*	53.3†
First Liberty bonds (1947), called 1935.....	1,933.2	9.2
Due after Jan. 1, 1939.....	5,215.9	24.8
Total	\$21,028.4	100.0
JUNE 30, 1934		
Due before Jan. 1, 1939.....	\$3,458.4‡	53.3†
First Liberty bonds (1947), called 1935.....	1,933.2	7.7
Due after Jan. 1, 1939.....	9,861.2	39.0
Total	\$25,252.8	100.0
JUNE 30, 1935		
Due before Jan. 1, 1939.....	\$10,000.8	38.3
First Liberty bonds (1947), called 1935.....		
Due after Jan. 1, 1939.....	16,093.9	61.7
Total	\$26,094.7	100.0

* Should read \$13,879.3.

† Should read 66.0.

‡ Should read \$13,458.4.

1934. As indicated in the footnote, it should read: Due before January 1, 1939, \$13,879.3; 66.0. The error can be discovered by noticing that the percentage distribution for 1933 as printed adds to 87.3 per cent instead of 100 per cent.

The second error occurs in the first line of the distribution of June 30, 1934, in which the amount is printed as \$3,458.4 instead of \$13,458.4. This error becomes obvious after the first one has been detected. If neither error were discovered, one would naturally read the table to mean that ten billion dollars of bonds had been retired or replaced by longer maturities during the fiscal year 1933-34. The corrected figures show that the reduction was only 421 million dollars. In this case there are more errors in the original data than in the per cents but they provide a check on each other. When ratios alone appear it is impossible to determine where the error lies. Except in the case of distributions totaling 100 per cent, the existence of error

would not be apparent unless the figure obviously disagreed with known conditions.

Referring again to Table 36, if the column giving the amount of the total debt were omitted, it would be very difficult by reading the per cents alone to comprehend the changes that have taken place. For example, the decline in the per cent of the debt due before 1939 from 66 per cent in 1933 to 53 per cent in 1934 might be ascribed to refunding operations during the year. The amounts show that the decline in per cent of early maturity was caused mainly by an increase in the total debt resulting from the addition of some four and one-half billion of longer term bonds maturing after 1939. In contrast to this the further decline from 53 per cent in 1934 to 38 per cent in 1935 of bonds maturing before 1939 was the joint result of a decline of about three and one-half billions in short-term maturities and an increase of over six billions in longer-term maturities. Thus we see how essential the amounts are in arriving at the proper interpretation of the changes in the per cents.

Sometimes additional relationships can be derived from a given set of data. If the original data are not shown the reader is prevented from working out ratios which may be of more interest to him than those selected by the author. Full presentation of the original data is therefore evidence of good faith on the part of the author. The reader is free to check every statement and to work out his own interpretation.

COMPARISONS BETWEEN RATIOS

Large and unwieldy figures are reduced to ratio form chiefly because comparisons between two or more such ratios can be easily interpreted. Many relationships that are entirely obscured in the original data can be brought out through the correct use of comparisons between ratios. In fact a comparison is so implicit whenever two or more related ratios are presented together that up to this point in the chapter it has been impossible to confine the discussion to single ratios and not to anticipate to some extent the relations that exist between the several ratios.

Kinds of Comparisons

These comparisons between ratios group themselves into two distinct types: (1) those between several ratios in a single series, all of which

have the same base, and (2) those between two or more separate ratios, in each of which the base is a different quantity or value. The first kind of comparison, with a very few exceptions, involves ratios between like items while in the second the ratios compared may be made up of like or unlike items. The methods used in the two kinds of comparison will be discussed separately.

Ratios on the Same Base.—There are two kinds of series in which several ratios are computed on the same base: (a) percentage distributions of parts of a total and (b) index numbers in which successive values in a time series are expressed as per cents of an earlier year or some other normal base period. The primary purpose in presenting either of these types of ratios in a series is to show the importance of each individual item in relation to the base of 100 per cent. Again it should be emphasized that the construction of either type of series presupposes the homogeneity of the data. In a percentage distribution the separate parts must comprise a unified whole and in a time series the successive values must be identically defined from period to period.

Percentage distributions: As was illustrated in Table 31, the expression of each part as a per cent of the total makes it easy to estimate the relative importance of each part. Stating the several per cents, along with the 100 per cent base, usually expresses the comparison sufficiently without any further computation. The difference between any two such ratios may also be expressed by subtracting one from the other provided the relation of this difference to the 100 per cent base is clearly indicated. In Table 31, for example, "Only 5 per cent more of the total mortgages was held by savings-and-loan associations than by the group next in importance, banks and trust companies." If a direct relation between any two items had been of main importance without reference to the total, the percentage distribution need not have been made. It would have been simpler to divide the two items of original data. However, if only a percentage distribution is available without accompanying data, dividing one per cent by the other will give the relation between the two items, since the identical denominators cancel out. To express the importance of savings-and-loan associations rela-

tive to banks and trust companies: $\frac{431.8}{1,437.3} \div \frac{359.2}{1,437.3}$ is equivalent to $\frac{431.8}{359.2}$; therefore dividing the two quotients, $30.1 \div 25.0$, gives the same result, 120 per cent. This operation, strictly speaking, should not be considered as a comparison of ratios, but merely as a substitu-

tion of the percentages for the original items in order to derive a simple ratio between them.

Index numbers of time series: The emphasis in this kind of series also is on the relation of each item to the base rather than on direct relations between any two of the several items. Since indexes are apt to be more readily available than the original data and comparisons may be wanted between two specific periods instead of comparing either one with the base period, there is frequent occasion for expressing comparisons between two individual index numbers. The procedure is exactly parallel to that already described for percentage distributions: the two values may be subtracted provided the difference is stated as a per cent of the base period; or, as has already been mentioned in the discussion of percentage increase or decrease, the second may be divided by the first, the same as would be done with the original items.

Ratios on Different Bases.—In the preceding section dealing with ratios on a common base the comparisons of such ratios were made in the same direction as the computation of the original ratios. That is, the numerator items differed from their base with respect to a single characteristic and the comparisons between the ratios were concerned with these same differences. When, for example, the characteristic was time in years, the various ratios computed on a base year were compared only with respect to their differences from this base year. No additional differences of any kind were introduced in the comparisons.

Comparisons between ratios that are on different bases involve more complex relationships since they are always cross-comparisons of the original ratios. The ratios compared may be made up of like items or of unlike items and in either case the comparisons will be concerned with differences in a characteristic that was not involved in the separate ratios.

Ratios of like items: Classification according to Characteristic: Since single ratios between like items are classified as time, space, or attribute, the comparisons between such ratios according to a second characteristic become a cross-classification of these three kinds of characteristics. Figure 33 presents this cross-classification with an example of each kind of comparison. Each of the three main groups of comparisons between ratios includes the three kinds of single ratios according to characteristic, with the exception of space comparisons of space ratios.

If the data for any of the examples in Figure 33 were set up in tabular form, it would be clear that since each ratio has a different base, there are no constant terms in the ratios that are being com-

FIGURE 33
CLASSIFICATION OF COMPARISONS BETWEEN RATIOS
OF LIKE ITEMS WITH EXAMPLES OF EACH

KIND OF COMPARISON BETWEEN RATIOS	KIND OF SIMPLE RATIO	EXAMPLE
Time	1. Time	Ratios of the amount of magazine advertising in December to the amount in November, compared for a period of years.
	2. Space	Ratios of the production of steel in Buffalo to the production in Cleveland, compared for each of 12 months of a given year.
	3. Attribute	Percentage distributions by economic classes of the total value of United States exports, compared annually for several years.
Space	4. Time	Ratios of the indexes of department-store sales for December of a given year to December of the preceding year, compared by Federal Reserve Districts.
	5. Space	(This combination is impossible because there cannot be cross-classification of spatial characteristics.)
	6. Attribute	Ratios of low-priced car sales to total passenger-car sales in a given year, compared by main geographic divisions of the United States.
Attribute	7. Time	The percentages of increase or decrease in value of United States exports over a period of years, the changes being compared by economic classes.
	8. Space	The percentage of sales in a given sales district to total sales by a wholesale hardware company during a given year, compared by types of product distributed by the firm.
	9. Attribute	Ratios of cash to installment sales in a given month, compared by departments, in a large department store.

pared. In each case, however, there are certain points in common between the ratios which allow valid comparisons to be made between them.

Tests of Comparability: Most important of these is the fact that the ratios are "like" in the same sense that original items are "like." That is, they are identically defined except for one characteristic. The numerators of the ratios being compared are like and their denomi-

nators are like, each set differing according to an identical classification which becomes the classification of the comparisons between the several ratios.

A table showing the ratios used in Example 1, Figure 33, and including the original data, demonstrates the method by which the "likeness" of numerators and of denominators in any case may be determined and the consequent possibility of drawing comparisons between the respective ratios. In Table 37, the original ratios are each between two months in a given year, December to November, and these ratios are compared for several years. The headings and the title, as explained by notes in the original source, indicate that in each month and from year to year throughout the period the magazine lineage is measured by an identical method. For each month the data represent from 80 to 85 per cent of all magazines in the United States, the reports being compiled regularly by *Printers' Ink*. The numerators in column 2 are therefore identically defined, and likewise the denominators in column 1. Since the several numerators differ from one another only according to the stub classification, time in years, and the corresponding denominators follow the same classification, the resulting ratios in column 3 differ only according to this same characteristic. Hence the ratios are "like" and it is justifiable to draw comparisons between them: for example, to observe that for every year during this period, except 1935, the lineage has been smaller in December than in November, the declines ranging from .2 per cent to 26.3 per cent.

If there has been any change in definition the data cannot be given under a single column heading without an explanation of the change,

TABLE 37
CHANGES IN MAGAZINE ADVERTISING LINEAGE
NOVEMBER TO DECEMBER, 1933-38 *
(thousands of lines)

YEAR	(1) NOVEMBER	(2) DECEMBER	(3) PER CENT DIFFERENCE (2)+(1)-100%
1933.....	1,899	1,791	— 5.7
1934.....	2,317	2,136	— 7.8
1935.....	2,201	2,334	+ 6.0
1936.....	2,736	2,731	— .2
1937.....	2,989	2,893	— 3.2
1938.....	2,251	1,658	—26.3

* United States Department of Commerce, *Survey of Current Business*, 1936 Supplement, p. 24; 1938 Supplement, p. 25; September, 1939, p. 23.

in a footnote, which will indicate that the data are not comparable. In Example 2, Figure 33, if the figures for steel production included only the city limits for the first few months and the metropolitan district for the remaining months, this fact would be noted and the resulting ratios could not be compared throughout the year.

The following example illustrates how an invalid comparison may be made due to a concealed change in definition which could have been discovered if the data had been tested by a tabular analysis similar to Table 37. This statement appeared in the editorial columns of a city newspaper:

CAMPUS-SHY CUPID

X.Y.Z. University is no matrimonial bureau, say alumni officials of that institution. A recent survey [1937] of the classes from 1928 to 1935, inclusive, shows that fewer than half of the coeds graduated from X.Y.Z. in those years have married. Of the alumnae who answered the questionnaire, the following percentages reported they had married: 1928, 54.3 per cent; 1929, 46.8; 1930, 42.5; 1931, 41.4; 1932, 34.7; 1933, 30.2; 1934, 20.3; 1935, 12.7.

In tabular form, the data might have been somewhat as follows:

YEAR OF GRADUATION	(1) NUMBER GRADUATING	(2)	(3)
		GRADUATES MARRIED BY 1937	
		Number	Percentage of Total
1928.....	300	163	54.3
1929.....	350	164	46.9
1930.....	400	170	42.5
1931.....	500	207	41.4
etc.			

A study of the stub and column headings shows that since all of the reports were made in the same year, 1937, there is no time comparison between the ratios. There was a time difference in the year of graduation, and this classification in the stub gives the appearance of a time comparison between the several rows. However, the heading of columns 2 and 3, "Married by 1937," indicates that in order to get the true definition of the terms of the ratios, the date of each class must be subtracted from this fixed date. The result becomes a classification by attribute—the number of years since graduation, or the successively shorter periods during which each class has been exposed to the "hazard" of marriage. The ratios do show that for each addi-

tional year since graduation a larger percentage of the members of any class will have married, but this fact scarcely requires proof. If a time comparison between ratios is desired, all the numerator items must be like in attribute, but the time of each ratio must be different. That is, each class must have been out of college for an equal length of time, the comparisons being made in successive years. If the question had been "Married within five years after graduation," then the resulting ratios would have been made in 1933, 1934, etc. Such data would probably give no clear evidence of either increase or decrease in the percentage of college women married.

Another example will show that any violation of the previously discussed principles governing the construction of individual ratios will destroy the significance of comparisons between them. One of these principles stated that the possibility of further subdivision of data must be kept in mind before combining in ratio form two general totals that appear to be "like" in definition. The invalidity of such a ratio may not be apparent until it is compared with one or more similarly constructed ratios, with results that are contrary to known facts. A research bureau collected from 30 manufacturing and wholesale concerns monthly data on the value of outstanding accounts and overdue accounts. From the totals of these 30 reports the ratio of overdue to outstanding accounts was computed as of the first of each month. However, when the July, 1937, ratio showed a noticeable increase over June, several of the concerns complained that the true situation was being misrepresented.

Table 38, giving the collected data and accompanying ratios, shows what happened as a result of combining diverse elements in a single total. The ratio of overdue to outstanding accounts for all 30 firms increased from 20.4 per cent to 22.7 per cent between June and July, 1937, due to a decrease in the denominator. This decrease in the total outstanding accounts can be charged entirely to the 6 food concerns. Their outstanding accounts showed a drop so great that it more than counteracted the slight increase shown by the other 24 concerns. The ratios of the food concerns were quite different from the other 24 in both months. The 6 food concerns were subsequently reported separately, thus eliminating dissatisfaction.

In this case, the numerator of each ratio consisted of 30 parts, each of which had a very definite relation to a corresponding part of the denominator. However, the 6 food concerns were so different from

TABLE 38

OVERDUE AND OUTSTANDING ACCOUNTS OF THIRTY CONCERNS
FOR JUNE AND JULY, 1937, AND THE RATIO OF OVERDUE TO
OUTSTANDING ACCOUNTS

TYPE OF CONCERN	MONTH	OUTSTANDING ACCOUNTS	OVERDUE ACCOUNTS	OVERDUE + OUTSTANDING (%)
6 food	June	\$173,901	\$ 61,780	35.5
	July	133,712	70,904	53.0
24 manufacturing	June	822,516	141,307	17.2
	July	836,410	149,212	17.8
30 combined	June	996,417	203,087	20.4
	July	970,122	220,116	22.7

the other 24 that when all were combined the result was a heterogeneous total that did not correctly represent either of the component groups. Consequently, neither the individual ratios between these totals, nor the comparisons between the ratios, could be given any definite interpretation. The only way to analyze such a situation is to assemble the data, part by part, and to study each individual relationship in order to discover which totals or subtotals should be used in deriving ratio comparisons.

Interpretation According to Kind of Relationship: The eight examples in Figure 33 afford illustrations of all kinds of ratios according to relationship: part to part, total to total, and part to total. Numbers 1 and 4 are comparisons between part-to-part time ratios, the first being ratios between two corresponding parts of the same year compared for several different years, and the fourth being ratios between corresponding parts of two different years compared in different areas. In Number 9 the ratios are between two component parts of total sales, compared according to a second attribute, the different departments of the store. The ratios in Number 2, compared for several successive periods, are between two cities in the United States and may therefore be regarded either as part-to-part or total-to-total relationships.

Numbers 6 and 8 are examples of comparisons between a set of part-to-total ratios and Table 38 illustrates the same kind of comparison. In this type of comparison there are as many separate denominators as there are ratios being compared, all identically *defined* but not identical in *amount*. Reduction to ratio form has in one sense placed all the numerators on a common base of 100 per cent, but it must be remembered that the numerators are now relative instead of

absolute amounts and hence are not subject to further computation with the same freedom as were the original data. Such cases need to be carefully distinguished from comparisons between the several parts within a single percentage distribution. In the discussion of ratios on the same base it was demonstrated that such ratios could be substituted for the original data in making computations. This is no longer true when the bases are different, since they do not cancel out when one ratio is divided by the other.

Comparisons between several series of ratios, each series having a single base, present the same situation as sets of single part-to-total ratios except that they are more complex and afford greater opportunities for misinterpretation. Numbers 3 and 7 from Figure 33 are examples respectively of comparisons between several percentage distributions and between several time series, both examples having been derived from the same set of original data.

Part A of Table 39 shows original data on United States exports cross-classified according to four years in time, and according to five subdivisions of the attribute, economic class. In Part B, several percentage distributions according to economic class are given, one distribution for each of the four years. These may be interpreted together, somewhat as follows: Finished manufactures comprise the largest share of United States exports and have been increasing in relative importance throughout the period, from 41.8 per cent in 1934 to 49.0 per cent in 1937. Crude materials, the group second in importance, and manufactured foodstuffs have each meanwhile formed a continually diminishing share of the total exports. Semi-manufactures and crude foodstuffs have maintained a fairly constant relative importance, except for the increase in semi-manufactures in 1937.

If a table such as this one were read too hastily, the horizontal rows of per cents might easily be mistaken for index numbers on some earlier base. With this misinterpretation the first row would indicate that exports of crude materials had decreased in value each succeeding year, which is of course not the case.

The actual indexes, which appear in Part C, and the per cents of increase or decrease in Part D, present an entirely different situation. The total value of United States exports has increased every year since 1934, the total in 1937 being 57 per cent greater than in 1934. A net increase for the four-year period has been reflected in every class of exports. The greatest and most consistent increases in proportion to

TABLE 39
UNITED STATES EXPORTS BY ECONOMIC CLASSES, 1934-37*

		A					B				
		MILLIONS OF DOLLARS					PERCENTAGE DISTRIBUTION BY ECONOMIC CLASSES				
ECONOMIC CLASSES		1934	1935	1936	1937		1934	1935	1936	1937	
Crude materials.....		653	683	668	722		31.1	30.4	27.6	21.9	
Crude foodstuffs.....		59	59	58	105		2.8	2.6	2.4	3.2	
Manufactured foodstuffs.....		168	157	144	178		8.0	7.0	5.9	5.4	
Semi-manufactures.....		342	350	395	678		16.3	15.6	16.3	20.6	
Finished manufactures.....		879	994	1,154	1,617		41.8	44.3	47.7	49.0	
Total		2,100	2,243	2,419	3,299		100.	100.	100.	100.	

		C					D				
		INDEX ON 1934 BASE					PERCENTAGE CHANGE SINCE 1934				
ECONOMIC CLASSES		1934	1935	1936	1937		1935	1936	1937		
Crude materials.....		100	105	102	111		5	2	11		
Crude foodstuffs.....		100	100	98	178		0	- 2	78		
Manufactured foodstuffs.....		100	93	86	106		- 7	-14	6		
Semi-manufactures.....		100	102	115	198		2	15	98		
Finished manufactures.....		100	113	131	184		13	31	84		
Total		100	107	115	157		7	15	57		

*Source: Statistical Abstract of the United States 1938, p. 450

1934 appeared in manufactures, semi-manufactures showing slightly greater proportionate gain than finished manufactures. Manufactured foodstuffs declined in 1935 and 1936, but in 1937 exceeded the base year value by 6 per cent. Crude foodstuffs showed no change in 1935, a slight slump in 1936, but in 1937 jumped to 178 per cent of the 1934 value. Crude materials rose to 105 in 1935, dropped to 102 the following year, and in 1937 amounted to 111 per cent of the base year value.

The interpretation of any set of parallel time series is full of hazards. Chief among these is a tendency to make cross-comparisons between the corresponding per cents in the various series as if they were absolute values instead of indicating that every change is in proportion to the base of its own series. Failure to read each change in relation to the base year may give the mistaken impression that the percentage of change refers to the year immediately preceding.

These two examples of comparisons between series of ratios have brought out two entirely different sets of relationships neither of which was very obvious from the original data. In neither case was it necessary to make any further computations in order to express the comparisons. When they are thus stated in general terms, or are implicit, these relationships are usually clear and can be readily understood because the corresponding ratios in the several series are all expressed as relatives of 100 per cent.

Ratios of unlike items: There is no difference from the foregoing method in dealing with comparisons between ratios made up of unlike items. They correspond exactly to total-to-total attribute ratios between like items and may be compared according to the same characteristics, time, space, or attribute. Figure 34 contains an example of each of the three classifications.

Since any single ratio between unlike items must have a numerator and denominator that are identically defined in time and space, the most common comparisons between several such ratios will be with respects to differences in either one of these two characteristics. In such cases the numerators of all the ratios will be classified according to either time or space, their respective denominators will have the same classification, and this will become the classification of the ratio comparisons, as in Examples 1 and 2 of Figure 34.

Ratios between unlike items are also sometimes compared with respect to an attribute, but since unlike items do not ordinarily possess

FIGURE 34

CLASSIFICATION OF COMPARISONS BETWEEN RATIOS OF UNLIKE ITEMS, WITH EXAMPLES OF EACH

KIND OF COMPARISONS BETWEEN RATIOS	EXAMPLE
Time	1 Amount of income tax paid per capita in the United States, compared annually over a period of years.
Space	2 Ratios of turnover per store for a retail clothing chain during a given year, compared for 10 different cities.
Attribute	3 Ratios of new orders received by a manufacturing concern to its shipments of finished goods during a given month, compared by type of goods.

a common set of attributes this is a situation less frequently found. In accounting practice, however, the unlike items may often be overlapping parts of two different attribute classifications of the same thing. These attributes have a relation to one another although they are not mutually exclusive. However, since both terms of such ratios are expressed in the same unit they are subject to a cross-classification according to some other attribute of the common unit. Number 3 is such an example: new orders and shipments are items that overlap, but both may be classified according to type of goods, and the ratios between them may be compared according to that attribute.

In other instances series of ratios between unlike items may occur in which all have the same denominator, the numerators being differentiated according to an attribute. Examples are: death rates in the United States for a given year according to specific causes of death; production of wheat per acre under identical conditions according to grade of seed sown; and the per capita consumption of beef, mutton, and pork in the United States. A number of such series, each on a single base, may be compared with respect to differences in any other characteristic common to both terms of the ratios. The deaths from various causes might be separated according to sex and the two sets of rates compared; the production of wheat from the different grades sown might be tested in several states to compare the effects of varying climatic conditions; and changing habits in per capita consumption of the three kinds of meat might be compared over a long period of years. The interpretation of comparisons between such series of ratios would be similar to the interpretation of series of ratios between like items that was illustrated in Table 39.

It will be recalled that ratios between unlike items are not expressed as per cents except in certain instances when both items have the same unit. The column heading of a set of ratios names the values below it as a certain number of the numerator unit per 1, 1,000, 100,000, etc., of a given denominator, as "per capita consumption, in pounds," "average individual income tax, in dollars," or "number of automobile deaths per 10,000,000 gallons of gasoline consumed." Consequently, the ratios in the table appear in the same unit as some of the original data, and it is more difficult to remember that they are relatives than when they appear in the form of per cents. If too many different combinations of these ratio relationships between unlike items are given in a single table, the resulting confusion may be as great as when too many kinds of per cents are used together. There is also the same necessity for guarding against changes in definition of numerators or denominators during the course of the comparisons. The danger of comparing heterogeneous totals must likewise be avoided, since unlike items that are selected as having a relation to each other may not have been given limitations sufficiently specific. Some of these more complex problems involved in comparisons between ratios will be dealt with in the next chapter.

Averaging Ratios

Averaging ratios is one method of comparing them. However, the principles involved are sufficiently different from those discussed in preceding pages to warrant a separate presentation. There are two rules that must be observed: (1) ratios cannot be averaged unless they are comparable in every respect; and (2) whenever ratios are averaged they must be weighted according to their relative importance.⁴

Comparability.—This principle carries over directly from the earlier discussion of comparability of the terms of individual ratios and comparability between ratios. The data must be homogeneous for the purpose at hand, and the numerators and denominators must retain the same definitions throughout.

In Table 40, column 3, the yield per acre of rye in the United States is given for three successive years, with the average yield for the three years combined. These data might well be considered as too general for some purposes, but for the purpose of comparing the

⁴ This is intended to apply primarily to the use of the arithmetic average. A geometric average is often computed without weights.

United States with other countries they are specific enough. The arrangement of the original data in columns 1 and 2 indicates that the definitions have remained constant: no change has occurred in the units of measure, acres, and bushels; the year in each case is the calendar year and not the crop year; and the total acreage and the total production have been secured from all the rye-producing states by methods of reporting that are essentially the same from year to year. Therefore the average yield for the three years, 11.4 bushels per acre, is valid from the point of view of comparability of the ratios from which it was computed.

Weighting.—The relative importance of ratios is determined by the values of their respective denominators. If ratios are weighted by their denominator values and then averaged, the result is the same as

TABLE 40
PRODUCTION, ACREAGE, AND YIELD PER ACRE OF RYE IN THE UNITED STATES, 1933-35*

YEAR	(1) PRODUCTION (1,000 bu.)	(2) ACRES HARVESTED (1,000 acres)	(3) YIELD PER ACRE (bu.)
1933	21,150	2,349	9.0
1934	16,045	1,942	8.3
1935	57,936	4,063	14.3
Total	95,131	8,354	11.4

* *Agricultural Statistics*, 1936, p. 25, United States Department of Agriculture.

a ratio between the sums of the original data of numerators and denominators. The latter method was used in Table 40: $\frac{95,131}{8,354} = 11.4$.

Whenever all the original data are available this is a much simpler procedure than weighting and averaging the individual ratios, as follows:

$$\begin{array}{r}
 9.0 \times 2,349 = 21,150 \\
 8.3 \times 1,942 = 16,045 \\
 14.3 \times 4,063 = 57,936 \\
 \hline
 8,354 \quad 95,131 \\
 \frac{95,131}{8,354} = 11.4 = \text{weighted average}
 \end{array}$$

Approximate relatives in round numbers may be used as weights instead of the actual denominator values, causing practically no difference in the average. The acreages for the three years are in the

approximate proportion of 7, 6, and 12. Using these weights and dividing by the sum of the weights gives the following result:

$$\begin{array}{r}
 9.0 \times 7 = 63.0 \\
 8.3 \times 6 = 49.8 \\
 14.3 \times 12 = 171.6 \\
 \hline
 \quad \quad 25 \quad 284.4 \\
 \hline
 \frac{284.4}{25} = 11.4 = \text{weighted average}
 \end{array}$$

It is therefore possible to average ratios for which no accompanying data are available provided that there is available instead a set of weights proportionate to the values of the actual denominators. If no information is at hand concerning the relative importance of the denominators, the ratios cannot be averaged.

The only case in which a simple average of ratios will give a correct result is when all the denominators are of equal importance, that is, when they are identical in value. However, this constitutes no exception to the rule that weighting is always necessary. Such an average is not *unweighted*, but all the ratios have been given *equal* weights.

Using some other set of weights in place of the denominators of the ratios will not give a valid average.⁵ Some other factors may appear to be of importance but the average will be distorted unless such factors are combined with the separate denominators before the original ratios are computed.

Table 41 shows in Part A an incorrect weighted average, and in Part B the same data averaged by the correct use of weights. The weights used in Part A are the numbers of workers at each wage rate. Obviously these are of importance in determining average wage increases, but if so they must be incorporated in the denominators of the original ratios. Instead of being merely wage *rate*, the numerator and denominator of each ratio should be *payroll*, that is, rate $\times$ number of employees, as in Part B.

Since the number of workers at each rate is assumed to be the same in 1936 as in 1926, the per cents of change in the payroll (Part B, column 6) are the same as the per cents of change in wage rates (Part A, column 3). The weights, however, are in different proportion since in Part B the numbers of workers have been multiplied by varying wage rates as of 1926. Reduced to a common base of 100 the weights in Part A would be 40, 32, and 28, but in Part B they would be 13,

⁵ An exception to this rule is explained in connection with Table 74, page 394.

TABLE 41
COMPUTATION OF AVERAGE OF PER CENTS OF INCREASE IN WAGES
PART A. (INCORRECT WEIGHTING)

(1) DENOMINATOR	(2) NUMERATOR	(3) RATIOS	(4) WEIGHTS	(5) WEIGHTED RATIOS
Weekly Wage Rates		Percentage of Increase* (2) ÷ (1) — 100%	Number of Employees	(3) × (4)
1926	1936			
\$10	\$12	20	25	500
50	80	60	20	1,200
40	44	10	18	180
			63	1,880 = +29.8%
				63

PART B. (CORRECT WEIGHTING)

(1)	(2)	(3)	(4) DENOMINATOR (Weights)	(5) NUMERATOR	(6) RATIOS	(7) WEIGHTED RATIOS
Weekly Wage Rates		Number of Employees	Total Payrolls		Percentage of Increase* (5) ÷ (4) — 100%	(6) × (4)
1926	1936		1926 (1) × (3)	1936 (2) × (3)		
\$10	\$12	25	\$ 250	\$ 300	20	5,000
50	80	20	1,000	1,600	60	60,000
40	44	18	720	792	10	7,200
			\$1,970	\$2,692		72,200 = +36.6%
						1,970

Total original data, $\frac{\$2,692}{\$1,970} = 136.6\% - 100\% = +36.6\%$

* Percentages of increase, decrease, or difference may be averaged, provided all are in the same direction. If both positive and negative differences occur in a series of ratios, the per cent relations must be used instead of percentage of difference in averaging. If the answer is wanted as a percentage of change or difference, subtract 100 per cent from the result, as the final step.

51, and 36. Consequently the average of the weighted ratios in Part B differs from that in Part A. The average in Part B is the correct one because it coincides with the ratio of the total original data, (total column 5 $\div$ total column 4, $- 100\%$). Similar proof cannot be applied to the method used in Part A. The total of column 2 divided by the total of column 1 minus 100 per cent equals $+ 36$ per cent whereas the weighted average is $+ 29.8$ per cent.

It might be argued that in an actual case the number of workers at each wage rate would not remain the same over a period of 10 years. In such cases the payrolls can be determined for each year and, if the ratios between them are weighted by the denominators as in Part B, the average will again agree with the ratio between the total original data. The usefulness of this average representing total payroll increase may be open to question. There might be no change at all in wage rates, or even in total number of employees, but a shift in the distribution of personnel at the various rates would cause an increase or decrease in the total payroll.⁶ An analysis of each rate separately would probably be of greater value. However, the average of changes in total payrolls as shown in Table 41 is of practical value in calculating the effects of planned payroll changes for a given number of employees.

PROBLEMS

- For each of the following pairs of items, compute the ratio and (1) state the relation in words; (2) give reasons for selecting the item you used as the base; (3) justify in terms of the text the number of units used in the base.

a)	Total tonnage of steel produced in 1938.....	18,692,957 gross tons
	Total tonnage of steel consumed by the automotive industry in 1938.....	3,155,906 gross tons
b)	Number of commercial banks in the United States reporting retail installment paper in their portfolios as of Dec. 30, 1939	10,382
	Amount of retail installment paper held.....	\$541,367,000
c)	Bales of cotton produced in the United States in 1938...	12,008,000
	Bales of cotton produced in Brazil in 1938.....	1,877,000
d)	The average weekly wages of steel workers in the United States was \$35.90 in 1929 and \$29.40 in 1939.	
e)	Population in United States registration area, 1930.....	118,560,800
	No. of deaths from diphtheria, 1930.....	5,822
- From any issue of the *World Almanac* select ratios illustrating each of the three sizes of base explained in the text. What is the relation between the numerator and the denominator of each ratio?

⁶ This question is developed in greater detail under **standardized ratios in the next chapter**, and weight bias is discussed in connection with **index numbers in chapter XIX**.

3. a) In 1938 the sales per salesperson in 8 department stores located in cities with population less than 20,000 was \$8,500; for 30 stores located in cities with population over 1,000,000 the corresponding figure was \$18,000.
- b) In 1935 the per capita sales of drugstores in Cleveland were 110 per cent of the sales in Detroit.
- c) In 1938 the average amount of each dollar of revenue set aside by Class I railroads to pay taxes was $9\frac{1}{2}$ cents.
- To what extent do these examples conform to the rules stated in the text for ratios between unlike items?

4. Given the following ratios:
- a) Dollars paid by industrial consumers of electricity to the number of kilowatt hours consumed during a single year in a given industrial area.
- b) Bank debits in New York City to bank debits in the rest of the United States during a given month.
- c) Average number of active spindles in cotton textile manufacturing in the New England area this year compared with the corresponding figure for last year.
- d) Time deposits of commercial banks to demand deposits of those banks as of a certain call date.
- e) Sales of Chevrolet passenger automobiles to sales of Ford passenger automobiles last year in the state of California.
- f) Imports of wheat into the United States from Canada to production of wheat in the United States for last year.
- g) Number of deaths caused by industrial accidents in New York State in January, 1940, to the number in January, 1941.

Classify these ratios in three ways: (1) like items or unlike items, (2) time, space, or attribute, (3) part-to-total, part-to part, or total-to-total.

5. From all the families with incomes of \$1,000 and over in the following table, what per cent have no automobiles? Show method of computation.

SELECTED FAMILIES IN PORTLAND, OREGON, WITH INCOMES OF \$1,000 AND OVER, 1933

INCOME GROUP	NUMBER OF FAMILIES	PERCENTAGE NOT HAVING AN AUTOMOBILE
\$1,000-1,499	1,426	34.2
1,500-1,999	1,068	25.0
2,000-2,999	701	16.7
3,000-4,999	300	13.0
5,000-6,999	45	2.2
7,000 and over.....	27	11.1

6. Given the following information concerning deposits and depositors in mutual savings banks and postal savings in 1932 (000 omitted for all figures):

STATES	MUTUAL SAVINGS BANKS		POSTAL SAVINGS	
	Deposits	Depositors	Deposits	Depositors
United States	\$10,040,000	12,700	\$783,000	1,540
New York	5,287,000	5,900	82,000	200
Minnesota	63,000	100	20,000	40

- a) Compute whatever ratios you consider necessary to compare methods of saving from the data given.
 - b) Write a statement of your findings.
7. Each student will be given one assignment for each part of this problem. Answer either (a), (b), (c), or (d), throughout.

A.

CIGARETTE CONSUMPTION IN THE UNITED STATES (Millions)

July, 1933	9,526	July, 1936	14,801
July, 1934	11,355	July, 1937	15,290
July, 1935	13,138		

Compared with the same month of the previous year, compute the percentage relation and percentage change.

- (a) 1934 with 1933 (c) 1936 with 1935
(b) 1935 with 1934 (d) 1937 with 1936

B. The following data show the amount of retail trade (in millions of dollars) in Buffalo in 1935 for 8 lines of trade (column 9 being "all others") and the total:

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	Total
54.9	37.8	22.9	26.0	8 6	6.8	17.3	6.3	24.8	205.4

Express the relation between the 2 items assigned to you, and the per cent which each is of the total retail sales.

- (a) Column 1 food, and column 7 eating and drinking places
- (b) Column 2 general merchandise, and column 8 drugstores
- (c) Column 3 apparel, and column 4 automotive
- (d) Column 5 furniture and household goods, and column 6 lumber, building, and hardware

C. Number of automobile fatalities and millions of gallons of gasoline consumed by motor vehicles in four states, 1934:

	N. Y.	TEXAS	IOWA	N. H.
No. of fatalities.....	2,903	1,579	531	104
Millions of gallons of gasoline...	1,501	875	404	71

Express the relation between gasoline consumption and deaths from automobile accidents in: (a) New York, (b) Texas, (c) Iowa, (d) New Hampshire.

8. Are there any of the ratios in the following that you cannot interpret? If so, explain what additional information is needed in order to draw valid conclusions from the ratios.

PRODUCTION OF MAPLE SUGAR AND SYRUP IN
THREE LEADING STATES, 1937 AND 1938

	AVERAGE TOTAL PRODUCT PER TREE				PERCENTAGE OF UNITED STATES PRODUCTION
	As Sugar (pounds)		As Syrup (gallons)		
	1937	1938	1937	1938	
Vermont	1.5	2.3	.19	.29	37.9
New York	1.8	1.7	.22	.21	25.7
Ohio	2.7	1.9	.34	.24	15.3
United States	1.8	2.0	.23	.25	100.0
Average price	.290	.283	1.60	1.61	
Percentage of 1925-29 average	101.0	98.6	77.7	78.2	

9. The following is quoted from the report of a tobacco manufacturer to the stockholders: "Government figures, with our own figures, prove that our Company obtained during the first ten months of 1940 . . . 59.93% of all the cigarette increase of the entire industry,"
- What additional data would be needed in order to determine the importance of this report?
10. Locate each of the three sets of ratios of problem 7 according to Figures 33 and 34, and state which of the simple ratios are part-to-part, part-to-total, or total-to-total relations.
11. Describe two separate methods of computing the average per cent living on farms in Table 32, page 241. State exactly what data would be needed for each computation.
12. a) Compute the percentage change in average value per contract of non-residential building contracts, from the following data:

1937		1938	
NUMBER	COST	NUMBER	COST
124,305	\$564,961,000	116,993	\$567,069,000

- b) With the following additional information discuss how much significance can be attached to your original ratio:

TYPE OF NON-RESIDENTIAL CONSTRUCTION	1937		1938	
	NUMBER	COST	NUMBER	COST
Private garages and sheds.....	96,514	\$27,423,000	91,147	\$23,798,000
Other private construction.....	26,711	413,072,000	24,497	402,762,000
Public works, public buildings and utilities	1,080	124,466,000	1,349	140,509,000

13. An industrial plant had the following number of employees at 2 different periods, with their respective total weekly payrolls.

TYPE OF EMPLOYEES	1925		1928	
	Aver. Number Employed	Aver. Total Weekly Payroll	Aver. Number Employed	Aver. Total Weekly Payroll
Administrative and clerical.	194	\$ 9,409.00	156	\$ 7,566.00
Skilled	320	14,630.40	235	10,744.20
Unskilled	608	13,254.40	731	15,935.80
Total	1,122	37,293.80	1,122	34,246.00

A union of unskilled workers used the foregoing totals to prove that there had been a decrease in wages of 8.2 per cent. Discuss.

REFERENCES

BAILEY, WILLIAM B., and CUMMINGS, JOHN, *Statistics*. Chicago: A. C. McClurg & Co., 1917.

Chapter VI lays down the pattern from which all subsequent work on ratios has been built.

BURGESS, ROBERT W., *Introduction to the Mathematics of Statistics*. Boston: Houghton Mifflin Co., 1927, chapter II.

JEROME, HARRY, *Statistical Method*. New York: Harper & Bros., 1924 chapter VIII.

CROXTON, FREDERICK E., and COWDEN, DUDLEY J., *Applied General Statistics*. New York: Prentice-Hall, Inc., 1939, chapter VII.

CROXTON, FREDERICK E., and COWDEN, DUDLEY J., *Practical Business Statistics*. New York: Prentice-Hall, Inc., 1934, chapter VII.

RIEGLER, ROBERT, *Elements of Business Statistics*. New York: D. Appleton-Century Co., 1927, chapter IX.

CHAPTER XII

APPLICATIONS OF RATIOS

THE basic principles and uses of ratios were discussed in the preceding chapter. Some more complicated cases of ratio analysis omitted there are presented in this chapter. The subjects included fall into three groups: (1) refined ratios and their application in standardized form; (2) compound ratios and the conditions for their interpretation; (3) some types of ratios used in particular fields of business.

REFINED RATIOS

General

In a refined ratio special care is exercised to define both numerator and denominator so as to exclude whatever extraneous factors tend to obscure the direct relationship between the two items. The advantage of the refined ratio lies in the opportunity of selecting in the denominator the item or items that are directly related to the numerator. Thus in vital statistics the ratio of measles cases to population under 16 years of age conveys much more information than the ratio of measles cases to total population. In the latter case a decline in the ratio over a period of years might be the result of an increase in the number over 65 years of age in the population, whereas the incidence of measles among those likely to contract the disease may have been unchanged.

The ratio of labor cost in a factory to total cost of manufacture is a useful figure, but the denominator contains two kinds of cost, fixed cost and variable cost. The ratio of labor cost (a variable cost) to total variable cost gives a figure which is more valuable to management in analyzing operations. In the same way safety departments of manufacturing plants get an over-all accident rate for the entire plant by taking the ratio of employees injured to number employed. This figure is of value only as a summary. The danger of accidents varies from one department to another both in frequency and severity. In a steel plant the tipping of a ladle in the furnace room is of infrequent occurrence but usually results fatally to workmen in the path of the hot metal. On

the other hand men engaged in piling steel rods for storage may be involved in injuries rather frequently but the injuries are seldom fatal. Similar contrasts can be made between departments in any type of manufacturing operation. To take account of these variations safety men compute the ratio, accidents to employees, for each department separately.

Both the numerator and denominator of each departmental ratio are refined further in order to facilitate the study of accidents. The resulting ratio, known as the accident severity rate, is the number of days' work lost through accidents¹ divided by the number of equivalent full-time days worked.² The rate may be expressed in any unit of time, per week, per month, or per year.

The study of deaths in automobile accidents furnishes another example of the use of refined ratios. Columns 1 and 2 of Table 42 contain the number of persons killed in automobile accidents and the population of the United States yearly from 1930 to 1938. The number of persons killed in automobile accidents per million population is given in column 5. The steady rise of this ratio from 1930 to 1937, interrupted only in 1932 and 1933, is the basic fact of the so-called "automobile menace." The marked decline of the ratio in 1938 appears to be the first evidence of abatement of the "menace."

That this conclusion is premature can be seen from further study of available information. The ratio of fatalities to population takes no account of the changed hazard resulting from increase in the number of automobiles on the highways. This factor is included in the ratio of fatalities to automobiles registered, shown in column 6. This ratio fluctuates irregularly from 1930 through 1933, rises to a high point in 1934, and except in 1937 has declined since that time. The 1938 ratio of 109 fatalities per hundred thousand automobiles registered is the lowest recorded during the nine-year period. This ratio indicates that the increased death rate after 1934 was not proportionate to the increased hazard represented by the number of automo-

¹ The number of days' work lost can be counted for temporary accidents but not for death, permanent disability, or permanent impairment. Even temporary disabilities such as the loss of a finger will lead to different numbers of days' work lost. Consequently standards have been established for each type of accident. Thus, according to one standard 6,000 days are allowed for death, 4,000 days for loss of an arm, 1,200 days for loss of a thumb and one finger, etc. Through the use of these standards, accident severity can be measured as between departments of a plant, between plants or between industries, as well as for different time periods. United States Bureau of Labor Statistics Bulletin No. 234, *The Safety Movement in the Iron and Steel Industry*, p. 278.

² The number of equivalent full-time days worked is obtained by dividing the total number of man-hours worked during a given period by the standard working hours per day

biles in use. In fact, the decline in fatalities per car since 1934 may be evidence of increased caution on the part of drivers.

But there is another factor to be considered. Cars are being driven more miles in recent years; hence the exposure to accident is increased. The ratio, number of deaths per 100,000,000 gallons of gasoline consumed each year, as shown in column 7, allows for the increased mileage of cars. If the average number of miles obtained per gallon of gasoline had remained fixed, then the number of gallons of gasoline consumed would have a constant relation to the number of miles cars were driven. The average number of miles per gallon has probably changed slightly during this nine-year period, but accurate information on this point is not available. Taking column 4 as a legitimate substitute for miles driven, we find that the ratio of deaths per 100,000,000 gallons of gasoline consumed has declined sharply from the high point reached in 1934. The ratios in columns 6 and 7 show that the increase in the automobile death rate through 1937 is due to the increased exposure of the population to the hazard of automobile accidents in spite of a decline since 1934 in deaths per motor vehicle registered and per gallon of gasoline consumed.

This conclusion does not in any way affect the desirability of a reduction in automobile fatalities. It does seem to indicate that, on the average, drivers maintain better control of cars in recent years and that

TABLE 42

FATALITIES IN AUTOMOBILE ACCIDENTS RELATED TO POPULATION,
MOTOR VEHICLES REGISTERED, AND GASOLINE CONSUMED, 1930-38

YEAR	(1)* PERSONS KILLED IN AUTO- MOBILE ACCIDENTS	(2)† POPULA- TION (000 omitted)	(3)‡ NO. OF MOTOR VEHICLES LICENSED (000 omitted)	(4)‡ NO. OF GALLONS OF GASOLINE CONSUMED (000,000 omitted)	(5) FATALITIES PER MILLION POPULA- TION (1) ÷ (2)	(6) FATALITIES PER HUNDRED THOUSAND MOTOR VEHICLES REGISTERED (1) ÷ (3)	(7) FATALITIES PER 100 MILLION GALLONS OF GASOLINE CONSUMED (1) ÷ (4)
1930....	32,540	123,091	26,545	14,751	264	123	221
1931....	33,346	124,113	25,814	15,408	269	129	216
1932....	29,196	124,974	24,115	14,250	234	121	205
1933....	31,078	125,770	23,874	14,224	247	130	218
1934....	35,769	126,626	24,952	15,292	282	143	234
1935....	36,023	127,521	26,231	16,264	282	137	221
1936....	37,500	128,429	28,166	17,855	292	133	210
1937....	40,300	129,257	29,705	19,218	312	136	210
1938....	32,000	130,215	29,486	19,610	246	109	163

* Figures compiled and published by the Travelers Insurance Company, Hartford, Connecticut.

† *Statistical Abstract*, 1938, p. 10.

‡ Mimeographed releases of the United States Department of Agriculture, Bureau of Public Roads. Gasoline consumption is by motor vehicles only.

further attempts to reduce the automobile accident toll will depend upon the compilation and study of more detailed information concerning the accidents that occur. This most likely means ratios subdivided for day and night accidents, accidents occurring on city streets and on open highways, accidents related to age of driver and age of motor vehicle, etc.

Standardized Ratios

Standardization of ratios consists in separating an over-all ratio into several mutually exclusive parts and computing a new combined ratio in the form of a weighted average of the several part ratios. The weights selected are a distribution of the denominators of the several part ratios according to some accepted standard, and these weights are held constant throughout any series of ratios that are being compared.

The use of standardized ratios originates in the field of vital statistics where standardized and corrected death rates, birth rates, etc., are employed in comparisons between different cities or sections of the country. The crude death rates of two cities may differ because of a difference in the age composition of the two populations although the death rate for each age group is identical in the two. The effect of variation in the age composition of different populations can be adjusted by the use of either standardized or corrected death rates.³

These methods can be transferred advantageously to the field of business ratios. In most cases the concept of the corrected rate rather than the standardized rate is applicable to business situations. What are known as "standardized ratios" in business applications are the equivalent of "corrected rates" in vital statistics. The business usage is defensible because the expression "corrected rates" is likely to convey the impression that the unstandardized rates contain errors. Such is, of course, not the case, but the standardization leads to more precision in the interpretation of results.

Department Store Example.—The experience of a department store will serve to illustrate the standardization of ratios. The officials were using the amount of the average sales check in four selected departments combined as a quick evidence of changes in business conditions.

³ Standardized and corrected rates are computed differently and lead to different results. Particular circumstances will determine which should be used in a given case. The details of both computation and interpretation are well presented in Raymond Pearl, *Medical Biometry and Statistics* (Philadelphia: W. B. Saunders Co., 1923), pp. 198-207.

The total dollar value of sales and the number of sales were subject to wide seasonal swings but the ratio of the two—the amount of the average sales check—exhibited little seasonal influence. In the past this ratio had proved to be a very sensitive indicator of approaching business depression or recovery. Table 43, column 3, shows the ratio for August and September, 1936.

TABLE 43
SALES, NUMBER OF SALES CHECKS, AND AMOUNT OF AVERAGE SALES CHECK
FOR A DEPARTMENT STORE IN AUGUST AND SEPTEMBER, 1936,
FOUR SELECTED DEPARTMENTS COMBINED

MONTH	(1) SALES	(2) NUMBER OF SALES CHECKS	(3) AMOUNT OF AVERAGE SALE
August	\$48,102	8,013	\$6.00
September	45,530	7,660	5.94

The declining tendency of sales, number of sales, and the amount of the average sales check caused considerable consternation because everyone expected September results to be above the August level. When the figures for the entire store became available a sizable expansion of sales was shown as well as an increase in the amount of the average sales check. The question then arose as to what had happened to the hitherto reliable preliminary indicator.

A study of the results by departments is given in Table 44. In Departments I and IV the percentage of decline in the number of sales checks exceeded the percentage of decline in sales, so that the amount of the average sales check increased. In Departments II and III the percentage of increase in number of sales checks was less than the percentage of increase in sales, which again resulted in an increase in the amount of the average check. It seemed strange then that the average check should have decreased for the four departments combined. The explanation lies in the fact that Departments II and III with small average sales checks had increased sales while Departments I and IV with somewhat larger sales checks showed great declines in sales. This combination of changes shifted the weights of the four departments so much that the combined ratio declined.

The change in the amount of the average sales check in Table 44 is dependent upon shifts in sales among the departments and changes in the average amount purchased per customer. Since the intention was to measure only the latter change, a means of eliminating the effect of

the former had to be devised. This was done by setting up a standard distribution of sales checks among the four departments and computing the amount of the average sales check each month for this standard distribution.

TABLE 44

SALES, NUMBER OF SALES CHECKS, AND AMOUNT OF THE AVERAGE SALES CHECK
FOR A DEPARTMENT STORE IN AUGUST AND SEPTEMBER, 1936,
FOUR SELECTED DEPARTMENTS COMPUTED SEPARATELY

Department	AUGUST			SEPTEMBER		
	(1) Monthly Sales	(2) No. of Sales Checks	(3) Amount of Average Sales Check (1) ÷ (2)	(4) Monthly Sales	(5) No. of Sales Checks	(6) Amount of Average Sales Check (4) ÷ (5)
I.....	\$10,416	1,010	\$10.31	\$ 4,293	380	\$11.30
II.....	9,622	1,595	6.03	11,889	1,862	6.39
III.....	21,840	4,621	4.73	26,400	5,112	5.16
IV.....	6,224	787	7.91	2,948	306	9.63
Total or average.....	\$48,102	8,013	\$ 6.00	\$45,530	7,660	\$ 5.94

TABLE 45

COMPUTATION OF THE AMOUNT OF THE AVERAGE SALES CHECK (STANDARDIZED)
FOR A DEPARTMENT STORE IN AUGUST AND SEPTEMBER, 1936

Department	STANDARD FIGURES		AUGUST		SEPTEMBER	
	(1) Standard Distribution of Sales Checks	(2) Standard Distribution of Sales Checks With Total of One (1) ÷ 8,182	(3) Amount of Average Sales Check	(4) August Standard- ized Sales (3) × (2)	(5) Amount of Average Sales Check	(6) September Standard- ized Sales (5) × (2)
I.....	845	.10328	\$10.31	1.065	\$11.30	1.167
II.....	1,729	.21132	6.03	1.274	6.39	1.350
III.....	4,615	.56404	4.73	2.668	5.16	2.910
IV.....	993	.12136	7.91	.960	9.63	1.169
Total or average.....	8,182	1.00000		5.97		6.60

The standard distribution of sales checks was obtained by taking the average monthly number of checks in each department for the year 1935. The selection of this standard was more or less arbitrary, but it approximated the actual distribution of sales checks among the four departments. The computation of the standardized average sales check is shown in Table 45. The standardized distribution of sales checks is given in column 1. The reduction of these figures so that their sum is unity is shown in column 2. The computation of the amount of

the standardized average sales check is shown in columns 3 and 4 for August and in columns 5 and 6 for September. The multiplications are indicated at the head of the columns. The average sales check increased from \$5.97 in August to \$6.60 in September after the effect of shifts in sales between departments had been adjusted.

There is some likelihood that the standard distribution of sales checks will lose its representativeness as time goes on. To avoid this contingency the distribution of sales checks for the most recent calendar year could be used as the standard. The results for long periods of time will not be comparable if a changing standard is used, but the purpose is merely to get a preliminary judgment from month to month; hence the long run situation is unimportant.

Labor Turnover Example.—Another point in our economic system at which standardized rates should be employed is in the measure of labor turnover. A measure commonly employed is known as the separation rate. It is the ratio of the number leaving employment during a period to the number on the payroll during the period. For example, the separation rate of a manufacturing plant would be measured as shown in Table 46

TABLE 46
COMPUTATION OF MONTHLY LABOR TURNOVER (CRUDE SEPARATION RATE)
OF A MANUFACTURING PLANT

MONTH	(1) NUMBER OF EMPLOYEES ON PAYROLL AT BEGINNING OF MONTH	(2) NUMBER OF EMPLOYEES LEAVING EMPLOYMENT DURING THE MONTH	(3) CRUDE SEPARATION RATE (per cent) (2) ÷ (1)
March	4,800	536	11.2
:	:	:	:
:	:	:	:
November	4,660	453	9.7

These figures show an appreciable decline in the separation rate from March to November. This is a crude rate which for some purposes may be satisfactory, but it would not be safe to conclude that there was greater labor stability in this plant in November. It is well known that out of a group of men hired at any time a certain number will dislike the work and leave within a few days, others will be found to be unsatisfactory and will be discharged after a brief trial. The remainder will continue working for longer periods and those who stay as long as one or two months are likely to remain with the company

for years. That is, the "employment mortality" decreases with increase in length of service. It would therefore be desirable to study the separation rate with the length of employment in the plant standardized.

The method of doing this is shown in Table 47. The stub shows the length of time employees had worked for the company as of March 1. Column 3 shows the number leaving employment during the month and column 4 shows the separation rate by length of time employed. These specific rates show a gradual decline as the length of time employed increases. The higher rate for those employed ten years or over reflects the separations due to retirement, disability, and death. The specific rates follow a course quite similar to that of the specific death rates of a population: that is, a high infant mortality rate, a gradual decline to past middle life, and then an increase at the upper ages.

Exactly what should be used as a standard would have to be determined for each plant separately.⁴ The use of the average distribution of employees for the last calendar year seemed best suited for the plant in question because it had a highly fluctuating labor force, varying between 2,500 and 6,000 within three years. The use of a five-year or even a three-year standard would have placed too much emphasis on a past situation and there was too much variation to attempt the selection of an arbitrary standard. The standard distribution of employees in column 1 was obtained by taking the average of figures similar to those in column 2 for the twelve months of the preceding calendar year. These one-year averages are expressed in the form of decimals totaling unity following the method shown in Table 45, column 2. This form saves one step in the computation. If the actual average distribution of employees were used in column 1, the entries in column 5 would be the number of separations that would have occurred in the standard distribution of employees at the specific rates for the actual distribution of employees in March. The total of the standardized separations divided by the total of the standard distribution would give the same standardized separation rate that was obtained by the method in the table. However, using column 1 gives the standardized separation rate directly as the total (.1021) of column 5. The same computations give the standardized separation rate for November (.1026), as shown in columns 6 to 9.

The summary at the bottom of the table shows that the decline in

⁴ If, however, a standardized rate were to be determined for several plants or an entire industry, the same standard would necessarily have to be used throughout.

TABLE 47
COMPUTATION OF THE MONTHLY SEPARATION RATE OF A MANUFACTURING PLANT USING A STANDARD DISTRIBUTION OF LENGTH OF SERVICE OF EMPLOYEES*

LENGTH OF SERVICE OF EMPLOYEES	MARCH					NOVEMBER			
	(1) Standard Distribu- tion of Employees with Total of One	(2) Number of Employees on Payroll at Beginning of Month	(3) Number Leaving Employ- ment during Month	(4) Separa- tion Rate (3) ÷ (2)	(5) Stand- ardized Separa- tion Rate (4) × (1)	(6) Number of Employees Payroll at Beginning of Month	(7) Number Leaving Employ- ment during Month	(8) Separa- tion Rate (7) ÷ (6)	(9) Stand- ardized Separa- tion Rate (8) × (1)
Less than 4 weeks....	.125	830	241	.2904	.0363	405	115	.2840	.0355
4 weeks to 8 weeks....	.024	100	24	.2400	.0058	115	28	.2435	.0058
8 weeks to 12 weeks..	.045	204	38	.1863	.0084	200	36	.1800	.0081
12 weeks to 26 weeks..	.080	325	41	.1262	.0101	485	61	.1258	.0101
26 weeks to 52 weeks..	.268	1,250	115	.0920	.0247	1,420	137	.0965	.0259
1 year to 2 years.....	.179	850	36	.0424	.0076	780	35	.0449	.0080
2 years to 5 years.....	.156	700	21	.0300	.0047	710	21	.0296	.0046
5 years to 10 years....	.078	341	10	.0293	.0023	360	11	.0306	.0024
10 years and over.....	.045	200	10	.0500	.0022	185	9	.0486	.0022
Total or average....	1.000	4,800	536	.1117	.1021	4,660	453	.0972	.1026

Crude separation rate..... March
Standardized separation rate..... November
9.72%
10.26%

* The data of this table are fictitious, but they are approximately those used by the author in an unpublished study of labor turnover.

the crude separation rate from March to November was entirely due to change in length of employment. The standardized rate indicates that there was essentially no change in the forces of labor unrest leading to separations.⁵

This example and the one dealing with sales checks of a department store show the advantage of standardizing ratios in order to separate the significant causes of an observed change in the crude ratios. The method has not been used much in the past by business statisticians, but a wider application can be expected in the future to meet the demand for more exact interpretation of the ratios used to measure changes in business operations.

COMPOUND RATIOS

When a ratio between two ratios is computed the result is known as a compound ratio. Such ratios require careful interpretation, because the computation is of the form,

$$\frac{\text{numerator}_1}{\text{denominator}_1} \div \frac{\text{numerator}_2}{\text{denominator}_2} = \text{compound ratio}$$

The result shown by the compound ratio may have arisen from changes in either or both numerators, either or both denominators or from changes in the numerator and denominator of either or both ratios. With so many possible combinations of causes, it is evident that misinterpretation may easily occur. The conservative conclusion, therefore, would be to eliminate compound ratios entirely. However, they have become an integral part of the business man's use of statistics; hence it is preferable to explain their valid use rather than to eliminate a technique which has considerable practical value.

Compound ratios can be divided into three groups: (1) those in which the denominators of the two simple ratios are stable; (2) those in which the simple ratio used as denominator is stable; and (3) those in which all of the constituents fluctuate. The latter type of ratio can be used only as a general indicator of changes while the first two can be interpreted specifically. The distinction between the three in both form and interpretation can be seen in the examples that follow.

⁵ Current information concerning labor turnover in manufacturing plants is released in mimeographed form monthly by the United States Bureau of Labor Statistics. The reports include data for "quits," "discharges," "lay-offs," "accessions," separation rates, and turnover rates. This release does not make use of the refined rates presented in the text.

Stabilized Denominators

In this type, one ratio is divided by another whose denominator is not identical with that of the first ratio but the difference is so slight that for practical purposes the compound ratio is valid. It is subject to the same interpretations and comparisons with other compound ratios in the same series that could be made if all the denominators were identical. For example, a study of changes in the loans and investments of member banks of the Federal Reserve System is presented in Table 48. The decrease in the proportion of assets invested in government securities can be seen in column 3 but column 4 provides additional information concerning the months in which the decline was most pronounced.

It might appear that the same facts could be shown equally well by ratios between the successive numerators, in column 2. Comparison of these simple ratios, listed in column 5, with the compound ratios in column 4 shows, however, that this is not the case. Column 5 indicates that the greatest decline in investments in government securities occurred from March to April, but column 4 indicates that the greatest decline in the ratio, investments in government securities to total loans

TABLE 48

TOTAL LOANS AND INVESTMENTS IN GOVERNMENT SECURITIES OF REPORTING MEMBER BANKS OF THE FEDERAL RESERVE SYSTEM IN 101 LEADING CITIES, JANUARY-JUNE, 1937

MONTH	(1) TOTAL LOANS AND INVESTMENTS* (000,000 omitted)	(2) INVESTMENTS IN GOVERNMENT SECURITIES* (000,000 omitted)	(3) GOVERNMENT SECURITIES AS A PER- CENTAGE OF TOTAL LOANS AND INVESTMENTS (2) ÷ (1)	(4) COMPOUND RATIO: PER CENT THAT EACH MONTH'S RATIO IN COLUMN 3 IS OF THE PRE- CEDING MONTH	(5) SIMPLE RATIO: PER CENT THAT EACH ITEM IN COLUMN 2 IS OF THE PRECEDING MONTH
January.	\$22,734	\$10,493	46.2		
February.	22,600	10,330	45.7	$\frac{45.7}{46.2} = 98.9$	98.4
March.	22,610	10,008	44.3	$\frac{44.3}{45.7} = 96.9$	96.9
April.	22,280	9,628	43.2	$\frac{43.2}{44.3} = 97.5$	96.2
May.	22,201	9,483	42.7	$\frac{42.7}{43.2} = 98.8$	98.5
June.	22,330	9,515	42.6	$\frac{42.6}{42.7} = 99.8$	100.3

* *Federal Reserve Bulletin*, September, 1937, p. 924.

and investments, occurred from February to March. Further, column 5 shows that investments in government securities increased from May to June, whereas column 4 shows that the ratio of investments in government securities to total loans and investments declined slightly from May to June, although the rate of decline was gradually tapering off. It is evident, therefore, that while both columns 4 and 5 provide valid comparisons the information contained in column 5 is in no sense a substitute for that contained in column 4.

Denominator Ratio Stable

This case arises where one ratio or set of ratios is used as a standard to which another ratio or set of ratios is to be compared. Trade and mercantile association secretaries frequently employ this form of ratio comparison in studying the distribution of costs of doing business of individual association members.

A wholesale grocers' association had 28 members. The members did not agree to divulge their actual sales or costs of operation, but each month they reported the per cent which each of the items listed in Table 49 was of the sales for the month. With the aid of an agreed-upon set of weights, proportionate to the sales of the individual members, the secretary of the association computed the average percentage distribution of the several items making up total sales. This distribution served as a standard to which each of the members could compare his own operations. Columns 1 and 2 of the table give the average

TABLE 49
PERCENTAGE DISTRIBUTION OF SALES OF WHOLESALE GROCERS INTO COSTS OF
DOING BUSINESS AND PROFIT; CONCERN NUMBER 10 COMPARED
WITH AVERAGE FOR 28 CONCERNS*

ITEMS	(1) PERCENTAGE DISTRIBUTION OF COSTS	(2) Member Concern No. 10	(3) PERCENTAGE VARIATION OF CONCERN No. 10 FROM THE AVERAGE (2) ÷ (1) — 100%
	28 Member Concerns	Member Concern No. 10	
Administrative expense	6.1	6.4	+ 4.9
Rent, interest, and insurance.....	1.7	4.2	+147.1
Selling expense	8.1	10.7	+ 32.1
Handling expense	2.0	2.4	+ 20.0
Delivery expense	4.5	6.1	+ 35.6
Cost of goods sold.....	76.3	69.3	— 9.2
Profit	1.3	.9	— 30.8
Total	100.0	100.0	

* Data taken from the files of the secretary of a wholesale grocers' association.

distribution and the distribution for one of the members, respectively. Comparison of the two columns shows the items for which the particular concern had either better or poorer than average results. However, the variation with regard to specific items is brought out more clearly by the compound ratios in column 3.

These ratios must be interpreted as showing not the percentage difference in absolute dollars by which each item of Concern Number 10 differed from the average but as dollar for dollar in proportion to total sales. That is, the actual dollar profit of this concern was not 31 per cent less than the average profit of the 28 together but for every dollar of profit that should have been made by Concern Number 10 according to the standard only 69 cents was realized, despite the fact that the cost of goods sold was 9 cents on the dollar less than the average. Similarly, this concern's selling expense was 32 cents per dollar greater than the standard, the handling expense was 20 cents greater, and the delivery expense was 36 cents greater.

The fact that selling, handling, and delivery expense exceeded the average was no cause for alarm. Concern Number 10 did a large proportion of its total business in fresh fruits and vegetables. It used a unique system of selling. Trucks were sent daily through all of the territory served by the concern. On the theory that truck drivers are not good salesmen, a salesman who had nothing to do with handling goods rode on each truck. His duty was to sell goods which the truck driver immediately delivered. This system added to the concern's selling costs and its delivery costs. The high handling costs arose from the nature of the goods. That is, car-load shipments of oranges, grapes, tomatoes, and similar things required extra handling on account of spoilage, the need for protection from cold weather, and the necessity of moving the goods quickly, so that the higher operating costs were to be expected. The lower cost of goods was likewise clear enough because perishable goods must be marked up a greater percentage on cost to provide for waste, spoilage, and the additional cost of quick sales. The small profit margin, however, was unsatisfactory and further investigation was undertaken.

The administrative expense was not much above the average but the investigation indicated that some economies might be made at that point. However, the great variation in the percentage of sales going for rent, interest, and taxes was a revelation to the owner. Further study showed that rent made up a great part of the total of this item.

This raised the question whether the concern could reduce the floor space used. The rent paid for the combined office and warehouse building in which most of the business was conducted seemed to be quite low, but two auxiliary warehouses in which bulky commodities such as potatoes were stored had been leased at high rentals. By rearranging the use of space in the main warehouse and reducing the inventories of these bulky commodities, the concern was able to abandon the use of both auxiliary buildings when the leases expired. The charge for rent was thereby greatly reduced without any increase in other costs. The saving was therefore reflected directly in increased profits.

This example shows what can be done through the interpretation of compound ratios when one of the two sets of ratios is used as a standard. The members of the association were able, through the use of the association secretary's report, to compare their own operations with a standard established under similar conditions. Yet no member of the association had divulged any information which he desired to be confidential.

Fluctuating Numerators and Denominators

Table 50 shows the computation of the change in the ratio of accounts receivable to sales from August to September. The 25 per cent increase in the ratio of receivables to sales in September computed as follows, $(50 \div 40) - 100\% = +25\%$, cannot be given any specific interpretation. Any information concerning the reasons for the increase must be obtained from a study of the original data from which the simple ratios were computed. Both the receivables and the sales increased but the receivables increased 50 per cent while the sales increased only 20 per cent, thus accounting for the increase in the ratio. Further study would be needed to discover why the receivables increased so much. The indications from the figures are that all of the increase in sales was credit business. If such proved to be the case management would have to consider the implications of a continuation of the same tendency in the future.

All of the interpretation of the table arose from study of the original data rather than the ratios. The ratio of receivables to sales would have indicated the change that took place just as well as the compound ratio. There is, consequently, no advantage in computing the compound ratio beyond that of further summarizing the information contained in the simple ratios.

TABLE 50
COMPUTATION OF PERCENTAGE CHANGE IN RATIO OF RECEIVABLES TO SALES
FROM AUGUST TO SEPTEMBER

MONTH	ACCOUNTS RECEIVABLE	SALES	RECEIVABLES SALES (per cent)	CHANGE FROM PRECEDING MONTH IN RATIO OF RECEIVABLES TO SALES (per cent)
August.....	\$20,000	\$50,000	40	
September	30,000	60,000	50	+25

There are many applications of compound ratios in the analysis of business conditions. Those uses in which the denominators of both the simple ratios or both terms of the denominator ratio are stable will lead to results which can be interpreted rather exactly. On the other hand those cases in which the numerators and denominators of both ratios vary freely are more difficult to interpret. Because of this difficulty it is better to avoid the chance of misinterpretation by using such compound ratios merely as a guide to information which may be obtained from study of the simple ratios involved.

EXAMPLES OF THE USE OF RATIOS IN BUSINESS

In this chapter and the preceding one considerable emphasis has been placed upon the widespread use of ratio analysis of business data, and numerous illustrations have been presented. Such usage ranges all the way from the simplest percentages to highly specialized ratios. Some of the latter are of particular interest because of the comprehensive analysis of business activity which results from their use. Four examples of specialized ratio analysis are presented here: (1) a railroad analysis; (2) a retail credit department analysis; (3) a department store analysis; and (4) a financial statement analysis.

Railroad Analysis

These ratios originate partly in the Interstate Commerce Commission and partly in the Association of American Railroads. They pertain to several phases of railroad operation as indicated in Table 51. A complete explanation of the ratios would require too much space in this book. Each ratio provides specific information concerning a certain phase of railroad operations. For example, row (c), the freight revenue per train-mile, and row (b), the passenger revenue per train-mile, con-

TABLE 51

RATIOS USED IN THE ANALYSIS OF RAILROAD OPERATIONS IN THE UNITED STATES
DATA FOR SELECTED YEARS SHOW THE TREND OF OPERATIONS*

RATIO	UNIT	1900	1910	1920	1930	1935
Freight Operations						
(a) Revenue freight per train-mile	ton	271	380	639	699	646
(b) Revenue ton-miles per mile of road	thousand ton-miles	735	1,071	1,597	1,481	1,185
(c) Revenue per train-mile.....	dollar	2.00	2.86	6.81	7.56	6.51
(d) Revenue per ton-mile.....	cent	.729	.753	1.069	1.074	1.000
Passenger Operations						
(e) Average journey per passenger	mile	27.8	33.5	37.3	38.0	41.3
(f) Av. passengers per train-mile.	person	41	56	80	49	47
(g) Revenue per passenger per mile	cent	2.00	1.94	2.76	2.72	1.94
(h) Revenue per train-mile.....	dollar	1.01	1.30	2.78	1.85	1.35
Finance						
(i) Investment per mile of road..	dollar	61,490	64,382	81,954	105,661	106,339
(j) Taxes per mile of road.....	dollar	233	431	1,262	1,519	1,062
(k) Operating income per mile of road	dollar	7,722	11,866	24,361	20,564	14,483

* Compiled from reports of the Interstate Commerce Commission and the Association of American Railroads.

tain ratios showing the income from each type of service per unit of transportation employed. The compound units, freight-train mile and passenger-train mile, are designed especially to measure the basic operation of railroading, the movement of a train over a mile of track.

The difference between the trends of freight and passenger traffic can be seen by studying the figures for these two ratios in the table. Freight revenue per train-mile increased each decade through 1930 but by 1935 had declined to \$6.51, a drop of 14 per cent from \$7.56, the high value of 1930. On the other hand passenger revenue per train-mile declined from \$2.78 in 1920 to \$1.85 in 1930, a drop of 34 per cent. A further decrease brought the 1935 revenue to about the 1910 level. The great decline in passenger revenue per train-mile can be understood by reference to "average passengers per train-mile," row (f). In order to render passenger service, the railroads were forced to run their trains even though the number of passengers per train-mile in 1935 was little greater than the number carried in 1900. In contrast the number of tons of freight per train-mile, row (a), rose steadily to a high value in 1930 and was only slightly smaller in 1935. The ability of the railroads to control the size and the frequency of operating freight trains has resulted in a much smaller decline in revenue per freight-train-mile than that experienced per passenger-train-mile.

The ratios of Table 51 are all of the same type and can be catalogued as follows:

1. They are ratios between unlike items expressed in different units, including a number of compound units.
2. All of the ratios are refined to some extent in both numerator and denominator, e.g., "revenue ton-miles" excluded non-revenue freight and empty cars transported; "miles of road" is defined to exclude yard and terminal multiple trackage, all side tracks and auxiliary tracks and duplicate main tracks (a four-track line 100 miles in length is counted as 100 miles of road).
3. No compound or standardized ratios are included.

Retail Credit Department Analysis

A set of ratios designed to measure the activities of a retail credit department was published at the University of Michigan.

This investigation was undertaken to learn and make generally available facts regarding the costs, problems, and performances of credit and accounts receivable departments in retail stores.

From the outset, the study was planned so that as many as possible of the resulting facts would take the form of typical figures which could be used as standards for appraising performance in individual stores.⁶

The ratios which were developed in the study have been in general use since then. They were the following:

Per cent of returns to net [credit] sales ⁷

Per cent of credit office payroll to net [credit] sales

Per cent of accounts receivable office payroll to net [credit] sales

Per cent of total payroll [credit department] to net [credit] sales

Per cent of losses from bad debts to net [credit] sales

Payroll cost per transaction in the accounts receivable office

Number of transactions handled yearly per accounts receivable office employee

Average yearly salary in the credit office, in the accounts receivable office, and in both offices combined

Per cent of collections to first-of-the-month outstandings

Per cent of credit sales to total store sales

These ratios are quite specialized but that is exactly what is needed in dealing with the peculiar type of work performed by the credit department. Their use enables credit managers to follow very closely

⁶ Carl N. Schmalz, "Operating Statistics for the Credit and Accounts Receivable Departments of Retail Stores 1927," *Michigan Business Studies*, Vol. I, No. 6 (June, 1928). Bureau of Business Research, School of Business Administration, University of Michigan, Ann Arbor, Michigan.

⁷ Net credit sales include both charge accounts and installment accounts. In the study the information was obtained separately for the two types of accounts, so that the ratio analysis could be made for the two separately or combined.

the efficiency of collections. Where credit information including volume and character of operations of the stores in an area is pooled in the hands of an association, ratio analysis of credit conditions for the entire area becomes possible. There is tremendous advantage to the individual stores in doing this, as it provides a standard with which each can compare its own results.

Analysis of Department Store Operations

The Bureau of Business Research of the Harvard University Graduate School of Business Administration publishes an annual report analyzing the operations of department stores. The report is based on information supplied by some 600 department and specialty stores located in cities in all parts of the country. A list of the ratios used in analyzing the information is shown in Table 52.

TABLE 52

RATIOS PERTAINING TO OPERATIONS OF 55 DEPARTMENT STORES IN THE UNITED STATES
HAVING SALES BETWEEN \$2,000,000 AND \$4,000,000 IN 1934*

RATIO	RESULT OBTAINED
Net gain	
Percentage of net sales.....	2.4%
Percentage of net worth.....	4.6%
Rate of stock turn (times a year)	
Based on beginning and ending inventories.....	4.4
Based on monthly inventories.....	3.8
Returns and allowances	
Percentage of gross sales.....	8.5%
Percentage of net sales.....	9.3%
Distribution of net sales	
Cash	42.0%
C.O.D.	5.3%
Charge	45.8%
Installment	6.9%
Sales per square foot of total space.....	\$13.70
Real estate cost per square foot of total space.....	\$0.69
Sales per employee.....	\$3,500
Losses from bad debts (percentage of charge sales).....	.95%

* Selected from Tables 19 and 20 of Carl N. Schmalz, "Operating Results of Department and Specialty Stores in 1934," *Bureau of Business Research Bulletin No. 96* (June, 1935), Boston: Graduate School of Business Administration, Harvard University, pp. 27 and 28.

The ratios are self-explanatory. They provide a variety of information concerning operations, all of which is essential to management. The changes in these ratios from year to year indicate trends in department-store operation. Likewise the figures for any year serve as a standard to which an individual store can compare its results. For example, a store of comparable size finding that its stock turnover was less than 1.0 per year would want to take steps to find what types of

merchandise were moving slowly and whether it would be feasible to reduce inventory or expand sales, or both. A store with 3 to 4 per cent bad debt losses would want to investigate conditions in the credit and accounts receivable departments. Similar variations in any of the ratios would lead management to investigate. Weak places in the organization can frequently be discovered and corrected through this type of analysis and comparison.

Analysis of Financial Statements

A method of rating a borrower as a credit risk has been developed by Alexander Wall.⁸ By the use of seven ratios a numerical basis is provided whereby bank executives are materially aided in determining the lines of credit that can be granted to customers. For a detailed explanation of the use of these ratios it would be best to read the reference given. We are interested mainly in the statistical application of ratios involved and that can be explained most readily with the aid of an example.

Table 53 shows the balance sheet and annual sales of a concern for a five-year period. From these annual reports the seven ratios of Table 54 can be computed for each year and for the five years combined. The exact source of the numerator and denominator of each ratio is indicated by the letters in parentheses. Thus the ratio of net worth to debt for the first year ($n \div m$) is obtained as follows: $(\$492,105 \div \$156,094) = 3.15$ or 315 per cent, and all of the other ratios are obtained by similar computations.

Some of these ratios have rather high values and in general very careful study is required to judge the concern's standing as set forth in Table 54. The next step, therefore, is to refer the individual ratios to a standard. The standard chosen is the set of average ratios for the five-year period shown in column 6 of Table 54. The compound ratios in columns A to E of Table 55 are obtained by dividing each of the ratios in columns 1 to 5 of Table 54 by the standard ratios in column 6 of Table 54.

These standardized ratios could then be averaged to obtain a credit index. But they are not all of equal importance. The next step, therefore, is to establish a set of weights that can be used to give the several ratios their proper importance in determining the credit index.

⁸ Alexander Wall and Raymond W. Duning, *Ratio Analysis of Financial Statements*. New York: Harper & Bros., 1928.

TABLE 53

ASSETS, LIABILITIES AND SALES OF A HYPOTHETICAL CONCERN ANNUALLY FOR A FIVE-YEAR PERIOD*

DATE	FIRST YEAR	SECOND YEAR	THIRD YEAR	FOURTH YEAR	FIFTH YEAR	TOTAL FOR FIVE YEARS
ASSETS						
(a) Cash	36,285	34,479	27,776	37,206	51,157	186,903
(b) Receivables	229,437	208,363	220,666	231,171	233,540	1,123,177
(c) Inventory	236,586	208,376	245,367	265,304	255,004	1,210,637
(d) Total current ...	502,308	451,218	493,809	533,681	539,701	2,520,717
(e) Plant and equipment	115,389	146,884	169,045	170,195	132,037	733,550
(f) Miscellaneous	30,502	28,940	39,708	75,219	65,184	239,553
(g) Total fixed	145,891	175,824	208,753	245,414	197,221	973,103
(h) Total.....	648,199	627,042	702,562	779,095	736,922	3,493,820
LIABILITIES						
(i) Payables	139,894	152,455	220,539	308,880	255,459	1,077,227
(j) Taxes	12,030	719	4,083	5,779	7,218	29,829
(k) Miscellaneous	4,170	4,183	3,195	10,261	29,843	51,652
(l) Total current ...	156,094	157,357	227,817	324,920	292,520	1,158,708
(m) Total debt	156,094	157,357	227,817	324,920	292,520	1,158,708
(n) Net worth	492,105	469,685	474,745	454,175	444,402	2,335,112
(o) Total.....	648,199	627,042	702,562	779,095	736,922	3,493,820
(p) Sales	2,590,000	2,590,000	2,660,910	3,068,720	3,262,808	14,172,438

* Wall and Duning, *op. cit.*, p. 297.

These weights are shown in column F of Table 55. They depend mainly on the accumulated judgment of Mr. Wall and his associates.⁹ The final step in the computation of the index is shown in columns G to K of Table 55. The weighted combined result appears in index form at the foot of columns G to K, Table 55. The concern shows a strong credit position in the first year, is somewhat weaker the second year, but still has a high index. Decided weakness appears in the last three years, but the last year demonstrates marked recuperative powers in the concern as compared with the fourth year.¹⁰

Summary

Many additional examples of the application of ratios in the analysis of business operations might be cited. Those presented here have been

⁹ The complete argument for the use of weights and the reasons for the particular set selected are given in the source, pp. 157-62.

¹⁰ Those interested in the interpretation of the credit index should read from the source, pp. 299-306.

TABLE 54
SIMPLE RATIOS COMPUTED FROM TABLE 53

DESCRIPTION OF RATIOS	VALUE OF RATIOS (per cents)					
	(1) 1st Year	(2) 2d Year	(3) 3d Year	(4) 4th Year	(5) 5th Year	(6) Average of Five Years
Current assets to current liabilities $(d) \div (l)$	322	287	217	164	185	218
Net worth to fixed assets $(n) \div (g)$	337	267	227	185	225	240
Net worth to debt $(n) \div (m)$	315	298	208	140	152	202
Sales to receivables $(p) \div (b)$	1,129	1,243	1,206	1,327	1,397	1,262
Sales to merchandise $(p) \div (c)$	1,095	1,243	1,084	1,157	1,280	1,171
Sales to fixed assets $(p) \div (g)$	1,775	1,473	1,275	1,250	1,654	1,456
Sales to net worth $(p) \div (n)$	526	551	560	676	734	607

TABLE 55
COMPUTATION OF THE CREDIT INDEX FOR EACH YEAR DERIVED FROM TABLE 54

COMPOUND RATIOS	RELATIVES OBTAINED FROM COLUMNS AS INDICATED, TABLE 54 (per cents)					WEIGHTED RELATIVES OBTAINED FROM COLUMNS AT LEFT, AS INDICATED					
	(A) 1st Yr. (1) $\div$ (6)	(B) 2d Yr. (2) $\div$ (6)	(C) 3d Yr. (3) $\div$ (6)	(D) 4th Yr. (4) $\div$ (6)	(E) 5th Yr. (5) $\div$ (6)	(F) Weights	(G) 1st Yr. (A) $\times$ (F)	(H) 2d Yr. (B) $\times$ (F)	(I) 3d Yr. (C) $\times$ (F)	(J) 4th Yr. (D) $\times$ (F)	(K) 5th Yr. (E) $\times$ (F)
Current	148	132	100	75	85	25	3,700	3,300	2,500	1,875	2,125
Worth to fixed..	140	111	95	77	94	15	2,100	1,665	1,425	1,155	1,410
Worth to debt..	156	148	103	69	75	25	3,900	3,700	2,575	1,725	1,875
Sales to rec.....	89	98	96	105	111	10	890	980	960	1,050	1,110
Sales to mdse...	94	106	93	99	109	10	940	1,060	990	990	1,090
Sales to fixed...	122	101	88	86	114	10	1,220	1,010	880	860	1,140
Sales to worth...	87	91	92	111	121	5	435	455	460	555	605
Totals	...	...	...	...	...	100	13,185	12,170	9,730	8,210	9,355
Index	...	...	...	...	...	...	132	122	97	82	94

chosen mainly to demonstrate the variety of practical uses of ratios. The reader will also note the extent to which the ratios are specialized to meet the needs of the particular type of business to which they are applied. These ratios have been developed as the result of long experience and careful study. They are powerful tools of analysis in the hands of skilled investigators. Their use by less well-trained persons, who are unable to adhere to the basic principles of ratios which underlie all of the more complicated applications, may easily lead to gross error in the interpretation of results. All of which points to a final observation. Ratios are perhaps the most widely used statistical technique, but it is also true that no other technique produces an equal amount of misuse in practice.

PROBLEMS

1. What refinement would you recommend in the denominator of each of these ratios?
 - a) Employees killed in train accidents to total number of employees of railroads.
 - b) The number unemployed in a community to the number of persons in the community.
 - c) The ratio of investments of banks to loans and investments to show the tendency over a period of years toward increased importance of investments in bank portfolios.
2. The following data were compiled from reports of the United States Bureau of Census and the American Medical Association:

STATE	PHYSICIANS PER 1,000,000 POPULATION IN THE U. S.	DEATH RATE PER 1,000 POPULATION IN THE U. S. REGISTRATION AREA (1927)
California	1,781	13.9
New York	1,669	12.3
Massachusetts	1,552	11.6
Maryland	1,501	13.2
Illinois	1,492	11.4
Pennsylvania	1,251	11.4
New Jersey	1,078	11.2
Wisconsin	1,056	10.1

- a) Why is the ratio of physicians to population highest in those states in which the death rate is highest?
- b) From the figures given compute the deaths per physician in each state. Is it true that the deaths per physician are highest where the ratio of physicians to population is lowest? Explain the result.

3. The following data* appear to indicate that the rate at which workers were re-employed in private industry from Relief and W.P.A. rolls was lower at the peak of industrial expansion in 1937 than in 1935.

Duration of Unemployment (number of months)	MARCH, 1935		MARCH, 1937	
	Number of Unemployed Workers on Relief and W.P.A.	Number of Workers Leaving Relief and W.P.A. Because of Private Employment	Number of Unemployed Workers on Relief and W.P.A.	Number of Workers Leaving Relief and W.P.A. Because of Private Employment
Less than 2....	110	15	104	12
2 to 4.....	152	23	131	16
4 to 6.....	204	20	147	14
6 to 9.....	219	9	222	10
9 to 12.....	255	7	260	9
12 to 18.....	392	7	415	8
18 to 24.....	477	7	573	10
24 to 36.....	515	5	521	6
36 to 60.....	404	4	397	4
60 and over....	651	5	583	5
Total	3,379	102	3,353	94

Over-all rate:

$$\text{March, 1935 } \frac{102}{3,379} = 3.02\%$$

$$\text{March, 1937 } \frac{94}{3,353} = 2.80\%$$

* The figures are the result of making necessary adjustments to insure comparability in the records available in a city which had about 28,000 gainfully employed according to the 1930 census.

- a) Compute the re-employment rate according to duration of unemployment for each period.
 - b) Do the results confirm the decline shown by the over-all rate? If not explain the discrepancy and devise a plan for the construction of comparable over-all rates.
4. State the three types of relation between the four elements of a compound ratio.
5. Given the following data concerning population and number of persons paying income tax in the United States:

	1929	1936
1. Population (estimated)	121,945,000	128,883,000
2. Number filing income tax returns.....	4,044,327	5,413,499
3. Percentage of population filing income tax returns..	3.32	4.20
4. Percentage increase in number filing returns in 1936		+26.51
5. Number of income tax returns by persons with net income over \$5,000.....	1,032,071	677,011
6. Percentage of those filing returns who had income over \$5,000	25.52	12.51
7. Percentage decrease in 1936 in number filing returns who had income over \$5,000.....		-50.98

- a) Are the per cents in rows 4 and 7 equally valid? Why or why not?
 - b) Do these two per cents show that income has increased in the middle brackets at the expense of larger incomes, i.e., that wealth is gradually becoming more evenly distributed? Discuss.
6. What basic error would be involved in averaging the ratios in column 3 of Table 49 of the text?
 7. The following data are computed from annual reports of the United States Steel Corporation:

	1929	1939
Output per man-hour.....	\$1.80	\$2.48
Average wage per man-hour.....	.674	.897

Labor leaders argue from these ratios that labor has not received a fair share of its increased productivity. What further evidence should be introduced before reaching a conclusion on this point?

8. Which of the credit department ratios on page 290 are favorable when they increase? Which are favorable when they decrease? Explain in each case.
9. Compute from Table 52 the change in net gain as a percentage of net sales, if all losses from bad debts were eliminated.
10. a) Explain the two methods of measuring inventory used in obtaining the stock turnover in Table 52.
b) Which of the two is more exact? Why?
11. The heading of the stub of Table 55 of the text is "Compound Ratios." Which of the three types of compound ratios are these?

REFERENCES

BURGESS, ROBERT W., *Mathematics of Statistics*. Boston: Houghton Mifflin Co., 1927.

Chapter III contains an explanation of the effect on the resulting ratios when weights are shifted.

RHODES, E. C., *Elementary Statistical Methods*. London: George Routledge and Sons, Ltd., 1933.

Chapter 5 gives an excellent exposition of the use of refined ratios.

WALL, ALEXANDER, and DUNING, RAYMOND W., *Ratio Analysis of Financial Statements*. New York: Harper & Bros., 1928.

Chapters V-XII deal specifically with ratio analysis, although the entire book should be read to understand the system of analyzing financial statements.

CHAPTER XIII

GRAPHS

INTRODUCTION

ANY representation of statistical values and relationships in pictorial or diagrammatic form is called a statistical graph. There are many forms in which such graphs are commonly used, and new and ingenious devices are constantly being worked out for the graphic presentation of statistical data.

Reasons for Using Statistical Graphs

Graphic methods have been developed to meet the needs of two major classes of users, the lay reader and the statistician.

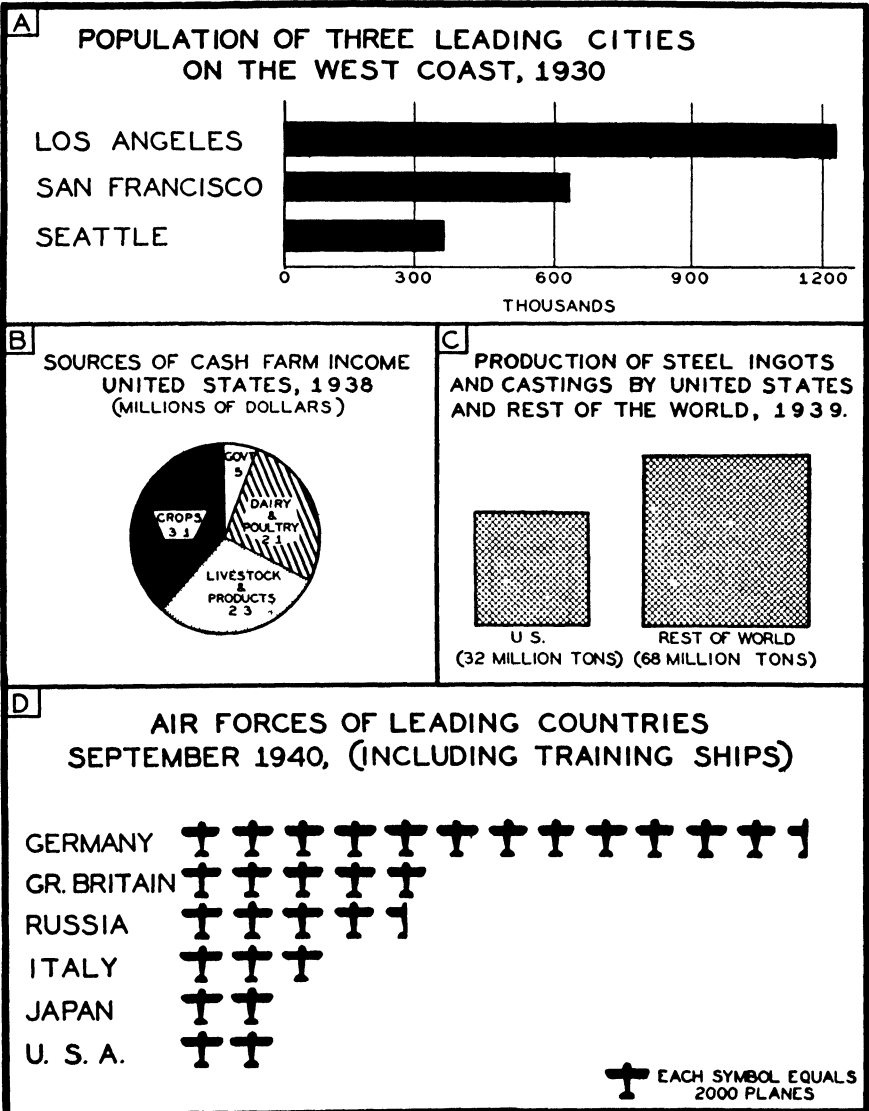
For the Lay Reader.—It is a well-recognized psychological principle that a direct visual concept such as color or size can be more quickly grasped and more readily understood than a set of numbers and table headings. A statistical table must first be read and then translated mentally into the actual concept of dollars of wages, bushels of wheat, etc. Although this is a process that is no more difficult than the reading of any other kind of printed material, the average reader is so frightened at the mere sight of a set of figures that he is likely to shy away from any table without even trying to see what it is about. In order to lure the attention of such readers, attractive graphs are an almost indispensable accompaniment to any popular exposition of statistical material.

For the Statistician.—Statisticians have also discovered that appropriate graphic methods have sometimes clarified relations that remained obscure even after careful study of the numerical data. Graphs of analysis have therefore become a tool of the statistician for his own benefit as well as a medium for explaining his final results to others.

Purposes Served by Graphic Methods

Standard methods can best be analyzed by reference to some examples of a few well-known kinds of statistical graphs. Four of these basic types are illustrated in Figure 35. The first, A, is a simple bar

FIGURE 35
TYPES OF GRAPHS



graph in which the *length* of a bar represents the population in each of the three cities. In B, the total cash farm income in the United States is represented by the complete *area* of a circle, and the proportion derived from each source of income is indicated by a sector of the circle, each sector shaded to distinguish it from the others.

Two squares are shown in C, the *areas* of which indicate the number of tons of steel produced by the United States and by the rest of the world in a given year. A *number* of equal-sized *symbols* in D indicates the approximate strength of the air forces of several countries in 1940.

From these examples the first statistical purpose of graphic methods can be observed.

To Show Numerical Values and Relationships.—In place of the actual figures that appear in the cells of a table, numerical *values* are represented by diagrams. These may consist of geometric forms such as the length of a line, the number of degrees in the sector of a circle, a square or other area, or they may contain a small number of symbols that scarcely need to be counted.

The numerical *relationships* between these values also can be grasped instantly without the necessity of reducing them to the form of ratios or other statistical measures. Referring again to Figure 35: In A the eye unconsciously estimates the lengths of the three bars, so that it is not even necessary to read the scale to perceive the relation of the populations of the three cities. In B the order of importance of the various sources of cash farm income and the proportion of each to the total can be seen without measurement. Likewise in D, the relative strength of the air forces of the several countries can be estimated by comparing the lengths of the several rows of symbols, without either counting the symbols or knowing how many units each symbol stands for. It is not so easy to compare the size of the two squares in C, and for this reason comparison by means of areas¹ is considered less effective than linear comparisons.

The simple types of graphic form in these illustrations show nothing except numerical values and relationships. No attempt has been made to introduce other characteristics of the data such as the spatial relationships between the cities in Part A, between the United States and the rest of the world in C, or between the various countries in D. However, the pictorial representation of non-numerical relationships such as these is the second major purpose of some statistical graphs.

To Show Non-numerical Relationships.—By the use of more complete and specialized types of graphs, time, space, or qualitative attributes can be presented in addition to numerical values. This is true

¹ Perspective drawings of three dimensional objects in which values are represented by volume are still more difficult to evaluate.

in spite of the fact that the number of possible methods for translating data into diagrammatic form is essentially limited to the few that have already been illustrated. There are not more than half a dozen methods altogether that can be used on a plane surface such as a page or a wall chart, but the number of ways in which dots, lines, areas, etc., may be combined in different arrangements has led to a seemingly endless variety of graphic types.

Every statistical table, except those in which the grouping is purely quantitative, contains in the stub or column headings a classification according to one of the non-numerical characteristics. In order to emphasize relationships between classes more realistically than is possible in tabular form, the data can be presented in certain specialized kinds of graphs.

Space: An alphabetical list of states conveys no picture whatever of actual geographical relationships; for instance, that Missouri is just across the river from Illinois; that Idaho, Nevada, and Utah form a group of mountain states; and that all the largest cities on the eastern seaboard fall within a radius of a few hundred miles. A statistical map in which certain numerical values pertaining to each state are depicted in their actual geographical arrangement lends itself to much more penetrating analysis of spatial relations.

Time: A column of dates may indicate successive years, while the adjacent column shows the volume of production in each year. This tabular form of presentation gives no such vivid concept of actual time movement and of the accompanying growth or decline in production as is afforded by a line whose movement across a graph one can follow along with the passage of the years.

Attributes: Qualitative attributes are difficult to represent graphically except by symbols. If black bars and shaded bars are used to represent male and female respectively, one has to read the key in order to have the slightest inkling of what the bars stand for. Rows of different symbols may be used instead of bars, as in Figure 35-D. This is one-third of the original graph. The middle section contained rows of warships representing the sea power of the six countries and the lower section contained rows of soldiers representing the armed forces of the six countries. In the three parts of the graph, therefore, unlike units rather than different attributes were being distinguished, but the same idea can be carried over to symbols representing attributes such as urban and rural, white and Negro, etc.

Construction of Graphs

Tests of a Good Graph.—Regardless of the particular purpose, every graph can be tested by one general criterion: The method of graphic construction is good if it produces an effective picture of important relationships and gives a true representation of statistical facts. Conversely, a method that results in an ineffective, uninteresting graph, or that distorts statistical facts, is bad.

Steps.—The steps in making a graph are the same as those followed unconsciously by a reader of the graph, but in reverse order. The reader's attention is first caught by the effectiveness, or attractiveness of a graph. If his interest is aroused, he then studies the graph and notes what information is presented through the various devices that have been used. Without obvious effort on his part, the facts and relationships which the graph was intended to emphasize will become clearly impressed upon his mind.

Interpret the data: For the statistician who plans and constructs the graph, however, this process of interpretation which is so easy for the reader becomes the first and most important problem. Before planning to illustrate any given set of data by means of a graph he must decide what relationship he intends to emphasize. He may wish to compare percentage changes in each of several commodities over a period of years, or their values in absolute amounts, or their percentage relation, either to each other or as parts of a total value in each year. The example of various ratio comparisons from a single set of primary data shown in Table 39, page 262, and the accompanying interpretations illustrate one of the possible initial steps of analysis that the statistician takes in extracting significant information from the data.

Choose the best graphic method: His next step is to determine which type of graph is best suited for presenting the relation he has determined is most important. There is often more than one way of picturing a statistical relationship, whether numerical or non-numerical, but years of testing have put the stamp of approval on certain methods as particularly effective for each purpose. These will be the subject of the balance of this chapter and the major part of chapters XIV and XV.

Draw the graph effectively: The final step is to plan the actual arrangement and draw the graph. The factors that contribute to artistic effect and accuracy are more or less common to all kinds of graphs, and will be considered in chapter XIV after the detailed discussion of

the simple types of graphs and some of the more complex graphs. It will be assumed throughout that the graphs are being prepared either for the business man who is not a trained statistician or for the general public. Graphs drawn for the use of statisticians should follow practically the same principles, although with greater emphasis on accuracy of detail and less on pictorial effect.

SIMPLE TYPES OF GRAPHS—METHODS AND PURPOSES OF EACH

In the ensuing discussion there is no intention of describing or even naming every kind of graph that has ever been used or that could be used. Only the commonly accepted types that have proved most successful in the presentation of statistical material² will be described in detail, and the major emphasis will be placed on the special purposes for which each is particularly well adapted.

Maps

The most obvious method for picturing spatial relationships is the map; it is more truly an actual "picture" of the facts than any other type of graph. There are many types of maps, ranging from the ordinary outline map which shows only geographical boundaries of land and water to so-called "distortion" maps at the other extreme. It is not the purpose of this section to describe all the possible maps that may be useful for various purposes, but only to point out the necessary features of maps that are definitely *statistical*. A map will be considered statistical only when quantitative relationships are represented by some pictorial device instead of in numerical form.

¹ *Location Maps*.—Outline maps form the background of statistical maps, and are also frequently used in business for non-statistical purposes. For example, a sales manager interested in knowing where each of his salesmen is working from day to day may move colored pins about on an outline map of the territory. But this is not a statistical map because *no numerical information is involved*.¹

Some location maps indicate non-numerical differences, such as states having certain types of liquor control or those in which a certain

² Business men use many types of diagrammatic representation that are adaptations of standard graphic methods. A variety of these, such as organization charts and Gantt charts are described in books on applied graphics. Space limitations have prevented the inclusion of a comprehensive presentation of these usages in a textbook on statistical method, although a few specialized forms appear in chapter XXVI.

gasoline corporation has retail outlets. Different colors or shades of cross-hatching are used in such maps, but unless the differences are in some way quantitative the map is not statistical, and its construction need not be governed by the rules given later for cross-hatched ratio maps. Many other location maps are found in print that present numerical information but are nevertheless not statistical. This is true of any map in which the quantities, ratios, etc., are simply inserted in each state or other subdivision in figures instead of being represented by shading or some other pictorial device. For most purposes such a map is harder to read than a table of the data, and no total visual impression is given of the spatial distribution of the numerical values.

A map in which evenly dotted areas signify the principal oil fields in the United States is also merely a location map. However, if each dot should be located at a point where there was production of 10,000 barrels of oil during a given period, the map would become statistical.

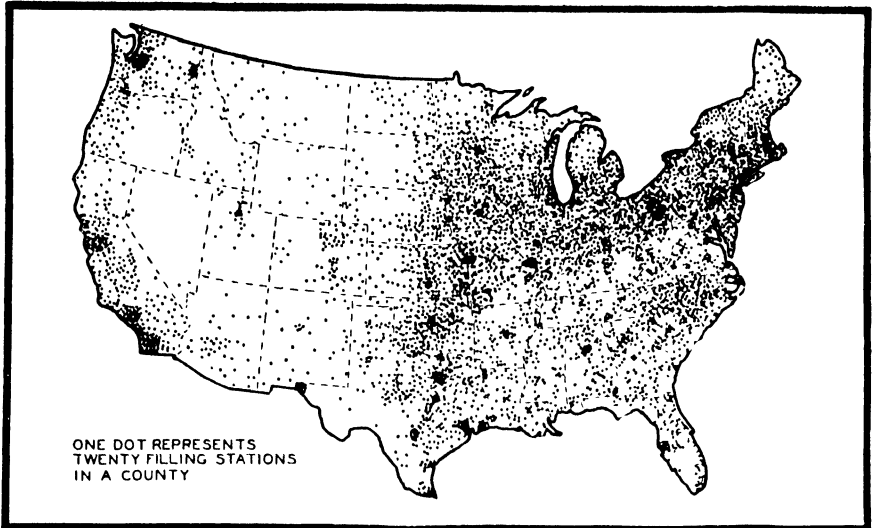
Dot Maps.—To show density: Such a map as the one last described is called a small-dot or point-dot map. The use of this type is illustrated in Figure 36-A, which shows the location of filling stations in the United States in 1935. Each dot represents 20 filling stations in a county, and, since the information was available, they have been clustered within actual county limits, even though county boundaries are not shown on the map. In some states the dots are so close together that it is impossible to count them, but a clear impression is gained of the general distribution of filling stations throughout the United States, that is, of the relative density of filling stations in each state and in various sections of the country. If too large an amount is represented by each dot, they will be so scattered that no great density is apparent in any subdivision; on the other hand if the unit is too small certain areas become entirely black. There is no intention that the dots should be countable in any subdivision but the unit must be so chosen that the effect of density will be clearly brought out.

To show quantity: 'If the actual totals in round numbers are wanted, large dots are used instead of small dots,' as illustrated in Figure 36-B. A certain effect of density is also provided by this map, but the dots are grouped in blocks by rows and columns, to facilitate counting, instead of being distributed throughout each state. The unit represented by each large dot is so chosen that no area will contain too many dots to count easily, and none will be too crowded to contain all the dots it should have. If dots must be drawn in the Atlantic Ocean to

FIGURE 36

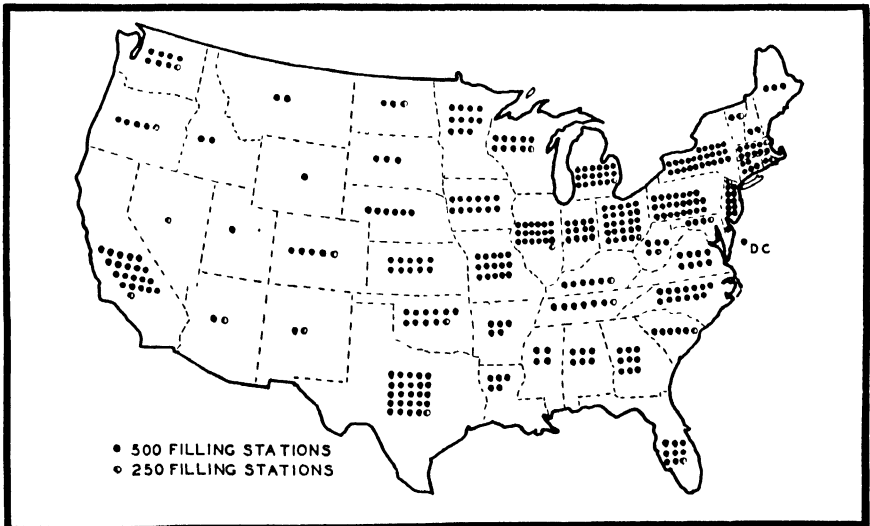
FILLING STATIONS IN THE UNITED STATES, 1935

A. Point-Dot Map Showing Density



Reproduced from *Market Research Series*, No. 18, United States Department of Commerce

B. Large-Dot Map Showing Number



Data from *Census of Business*, 1935

represent Rhode Island, Connecticut, Delaware, and other small states, the true spatial relationship is distorted.

It should be noted that, except for the last dot in each state, each dot represents an exact amount. In this example, each solid dot stands for exactly 500 filling stations, except that the last solid dot stands for the round number 375-625 when no half dot has been added. When a half dot has been added, each whole dot represents exactly 500, but the half dot may represent any number from 125-375. There will never be more than one partial dot in any given area. Partial dots are sometimes subdivided into quarters; i.e.; $\frac{3}{4}$ black indicates $\frac{3}{4}$ of the total unit, $\frac{1}{4}$ black indicates $\frac{1}{4}$ of the total unit, and an empty circle may be used for less than $\frac{1}{4}$ of the unit amount. However, this practice tends toward greater precision than is necessary in a graph.

The use of large dots is similar to the method of equal-sized symbols illustrated in Figure 35-D, and practically the same effect would be secured if symbols of gasoline pumps were used in Figure 36-B instead of large dots. Other forms of large dot maps can be found in print, in which the large dots are not all uniform. For example, instead of 5 equal-sized dots to represent 2,500 filling stations, a single dot 5 times as large might be used. This method involves the difficulty of estimating the relative areas of circles. In other cases there may be an attempt to show two or more different sets of information on the same map by means of large dots that are equal-sized but in several colors or shadings. This method is likely to result in a confused picture of spatial relationships instead of one that stands out clearly. Any deviation, therefore, from the solid large dot as illustrated in Figure 36-B is usually unsatisfactory for the purpose of showing geographical distribution of quantities.

Ratio Maps.—Although both small- and large-dot maps usually represent absolute quantities, there is an implied ratio even in these, because the quantities are distributed with relation to the *area* of each subdivision. This is particularly true of the small-dot or density map. The actual space covered by the area of each state is, in a sense, the denominator, and the number of point dots in each is the numerator. The resulting effect of density becomes a pictorial representation of the ratio, number (of some unit) to area. In Figure 36-A, for example, density increases either when the denominator (state area) is decreased or when the numerator (number of dots or filling stations) is increased.

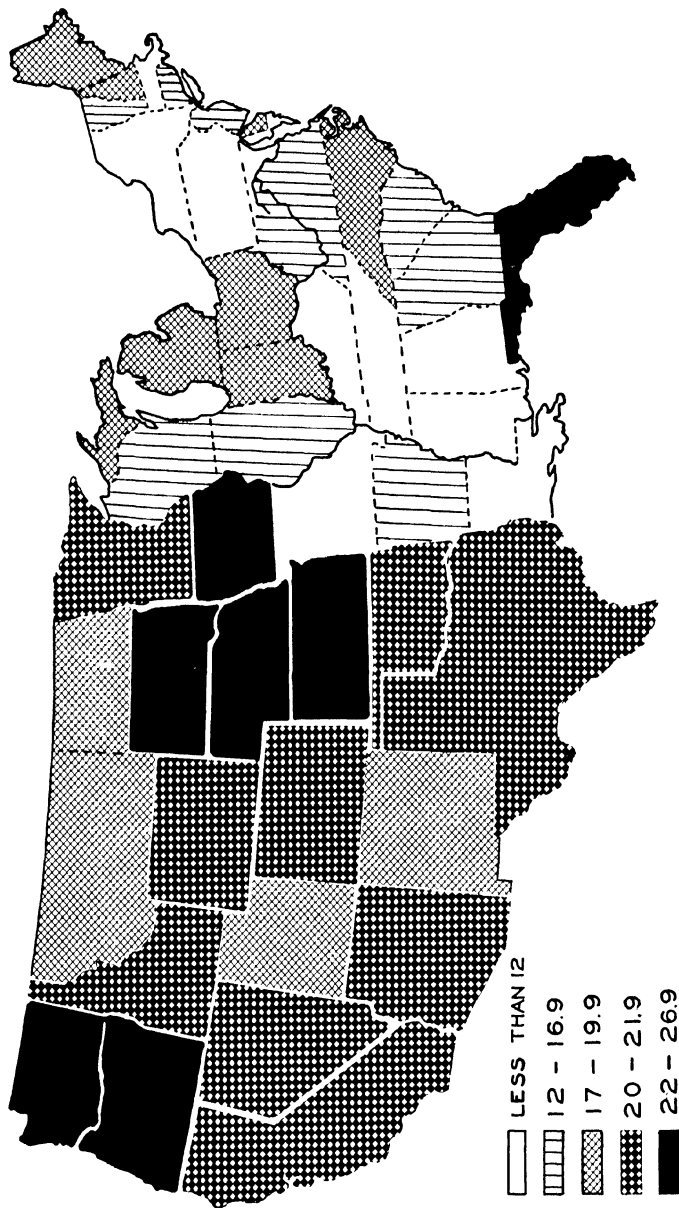
However, the purpose of a statistical map is often to show ratios in which the denominators are some values other than areas; for example, retail sales data might be used to show percentage of change over the preceding year, or sales per capita. Some pictorial device other than dots must therefore be used to summarize these ratios for spatial comparisons between the different sections of the map. The most usual method is by shading or cross-hatching, as illustrated in Figure 37.

Principles of cross-hatching: The types shown in the example are by no means the only possible kinds of cross-hatching, but they illustrate the principles involved. (1) The smallest ratio, i.e., the least density of occurrence of the characteristic, is best represented by white, and the largest by solid black, since these afford the greatest range of contrast. (2) The degrees of density should be in unmistakably ascending order from white to black. (3) This gradation is accomplished by increasing the heaviness of the lines, decreasing the space between them, or both; crossing them and filling in the intervening spaces are the next steps. Changing only the *direction* of the lines does not have any effect on the density of appearance. Alternate heavy black and white stripes may or may not appear darker than crossed or plaid effects and should therefore not be used in combination with them. The use of dots is undesirable for two reasons: it is often difficult to compare the density effect with that of lines, and there is danger of confusion with the point-dot type of map.

Interpretation of a cross-hatched map: In Figure 37, the same data that were used in Figure 36-B are expressed as ratios to the population of each state.⁸ There is no unmistakable impression to be gained from a single glance at this map, except that in the states west of the Mississippi River the ratio of filling stations to population tends to be higher than in the east. It is highest in the central farm states, in Washington and Oregon, and in Florida. The last named is most easily explained since the presence of many out-of-state cars produces a need for more filling stations than the native population would re-

⁸ Note that the purpose of this map is to show the density of filling stations with relation to the population. In order to avoid fractions of filling stations, the denominator has been expressed as "per 10,000 persons" instead of "per capita." An alternative method for stating the ratios as whole numbers would be to invert them, using "number of persons per filling station." If this were done, 400 persons per station would mean greater density of filling stations than 1,000 persons per station. The *lower* ratio would then more properly be represented by dark shading and the higher one by light, resulting in the same total effect as in Figure 37.

FIGURE 37
CROSS-HATCHED RATIO MAP: FILLING STATIONS PER 10,000 PERSONS IN THE UNITED STATES, 1935



quire. A study of the degree of car ownership and the miles of surfaced roads per capita in each state would aid in interpreting these ratios, but even a knowledge of these total state data is not enough. Neither will explain why such rural and relatively poor states as Mississippi, Alabama, Georgia, Tennessee, and Kentucky should fall in the same category as wealthy, industrial, and urban New York, Pennsylvania, and Massachusetts.

It takes careful study of the situation to realize that there is one pertinent factor common to these two kinds of states: a large proportion of the population does not own cars, and most of the car owners drive only short distances daily. This is the reverse of conditions in the comfortably prosperous farm belt and also in the large western states where distances are great. In the largest cities, such as New York, Chicago, Philadelphia, Baltimore, and Boston, traffic congestion is so great that a small proportion of families owns cars in comparison with small town and rural families in the same income groups. Thus the state ratio of filling stations to population will be reduced accordingly in New York, Illinois, Pennsylvania, Maryland, and Massachusetts. In the poorer regions of the south, comprising the greater part of the white and lightly hatched states, incomes are in general too small to lead to car ownership. North Carolina with its growing industries and improved roads is an exception in this area.

Because of these different factors that cause both numerators and denominators to fluctuate in the ratios depicted on this map, the picture presented is not as clear cut as can sometimes be achieved when the rates or other ratios pictured are definitely affected by a single condition of topography, climate, transportation facilities, concentration of population and industry, etc. In any case the planning of an effective cross-hatched map requires in a marked degree the co-ordination between artistic ability and statistical judgment which is stressed in the last part of chapter XIV. The rates, prices, per cents, or other ratios should be so grouped that if there are any significant spatial relationships they will stand out in the finished product. The choice of the right groupings is a very important factor in achieving this end.

Number and width of size groups: The first step in making a cross-hatched map is to work out the individual per cents or other ratios that are to be shown. These are next arrayed in order of size and studied

for any logical grouping that may be seen. If no natural dividing points are obvious, the items should be grouped so that an approximately equal number will fall in each category. In Figure 37, the number of states in each group is as nearly equal as the data permit—10, 11, 12, 9, and 7—although the size groups are of uneven width—under 12, 12–17, 17–20, 20–22, and 22–27. This is a departure from the rules that will be noted later in chapter XV for the presentation of frequency distribution tables and graphs.⁴

The number of groups has in this case been limited to five. More than six or seven kinds of cross-hatching are seldom effective on a map, and, in order to emphasize the contrasts, it may be desirable to have as few as three magnitude classes.

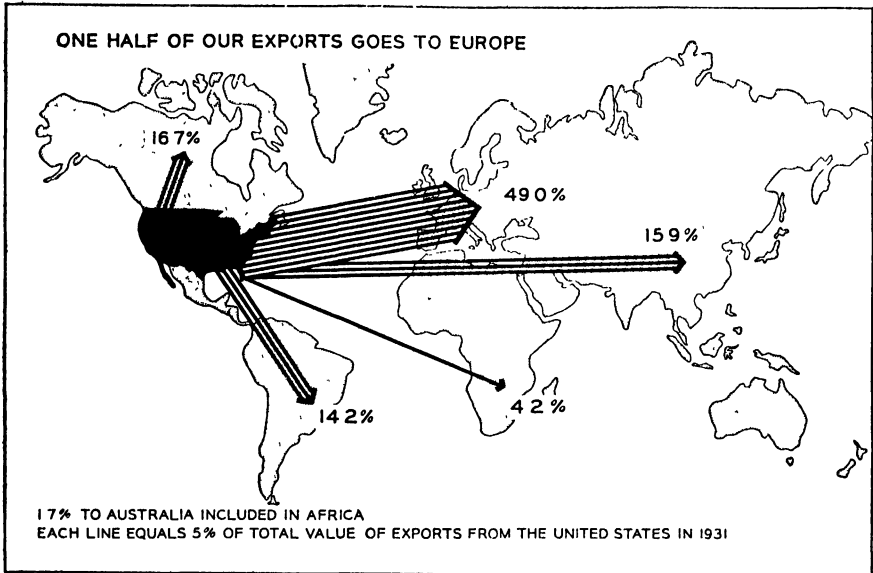
Special varieties of ratio maps: Certain conditions that are not confined within political boundaries may also be indicated by areas of cross-hatching. This kind of map is used chiefly to show belts of rainfall, crop conditions, etc., all of which are rates, per cents of normal, or other ratios grouped in class intervals. Non-statistical terms, as "good," "fair," and "poor" are sometimes used instead of a numerical measure for indicating crop, weather, or business conditions, but any such classes should be based upon numerical standards generally understood or defined in accompanying notes or text, and the rules stated above for ratio maps should govern the plan of shading.

Flow Maps.—These maps use a device not previously named, that is, numerical values are represented by the *width* of lines instead of by their *length*. The direction of the lines adds a non-numerical spatial relationship. This method has been found valuable chiefly in studies of traffic density. The same idea is followed in connecting areas of supply with areas of distribution. This is illustrated in Figure 38 showing the flow of exports from the United States to Canada and to other continents. Each line represents 5 per cent of exports, and the width of the combined lines indicates the proportion going to each area.

An alternative method utilizes circles of varying sizes to express the quantities at the point of supply, with arrows indicating the direction of distribution. Either method may be employed in simple diagrammatic form instead of on a map as background.

⁴ The entirely different graphic presentation of similar data in a frequency diagram requires class intervals of equal widths, whereas the number of items varies from class to class. See Bruce D. Mudgett, *Statistical Tables and Graphs* (Boston: 1930), Houghton Mifflin Co., pp. 179, 187–89.

FIGURE 38
FLOW MAP: UNITED STATES EXPORTS, 1931



Circle Graphs

For certain purposes a circle graph is superior to bars or other linear measures. However, by its very nature it is adapted to but few usages.

Parts of a Total.—The circle is a classic symbol of unity; hence it is ideal for representing total values. When divided into sectors, as in Figure 35-B, the same visual effect is given whether these values are expressed in per cents or as actual amounts. The actual total of cash farm income was \$8,000,000,000; the part contributed by crops was nearly \$3,200,000,000 or 40 per cent of the total, but whichever way it is designated its sector will cover about 144 degrees, or $\frac{2}{5}$ of the circle. Furthermore, since the angle measured by a given number of degrees is the same regardless of the size of the circle, it is possible to use two or more circles whose areas⁵ are proportional to compare the absolute amounts of several totals, yet the number of degrees in the corresponding sectors will also be comparable. This advantage is one not possessed either by linear bars or rectangular areas in representing parts of several totals. However, there is a certain tendency

⁵ The areas of circles are in proportion to the squares of their radii.

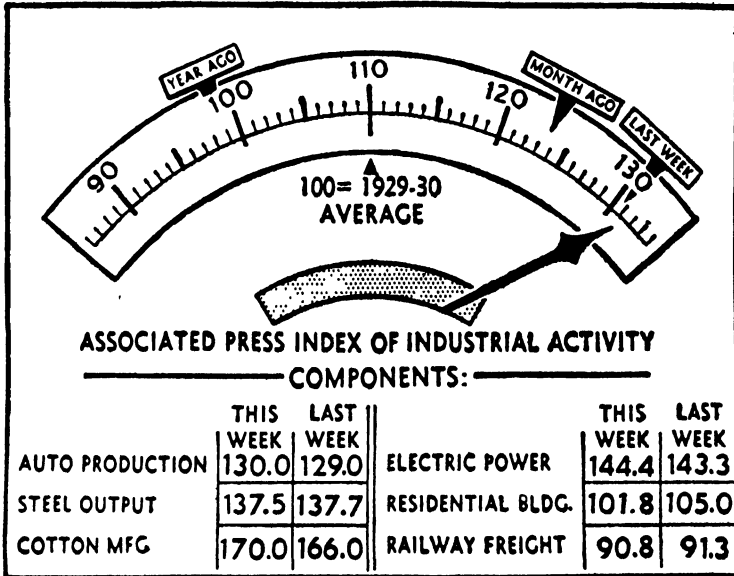
for the eye to measure the entire area of a sector instead of its angle. Therefore, if the emphasis is on comparison of the proportions of corresponding parts rather than on difference in total magnitudes, it is probably better to use equal-sized circles.

If the graph includes two or more circles, there should be a starting-point common to all of them, usually the radius extending upward from the center, although other quarter positions are also permissible. In each circle the various sectors should follow the same order, clockwise, around to the starting-point. The cross-hatching of sectors usually indicates different attributes or geographical divisions rather than quantitative differences, as in the case of cross-hatched maps; hence an ascending scale of density is not necessary. A better contrast is achieved if dark sectors alternate with light ones, as in Figure 35-B. The kinds of cross-hatching to be used therefore need not be definitely distinguished in degrees of density so long as no two sectors look too much alike. Dots are permissible and concentric curved lines are also used. Diagonal lines, however, should follow the same direction on the page in each circle, regardless of the direction of the radii in the sector. A key to the cross-hatching may be used instead of printing the legends in each sector, particularly when there are two or more circles.

Dial Indexes.—During the depression decade when practically every measure of business conditions was below normal, it became customary to use circle diagrams to show percentage of normal, the entire circle representing 100 per cent. These graphs were easy to understand but basically incorrect. In this case 100 per cent is not a total or maximum value but merely indicates a normal or average condition; it is possible for any index to rise above 100, whereas a circle can never represent more than 100 per cent. After some attempts to show an extra "piece of pie" on top of the whole, or bulging out at one side, the dial form of index was developed. In these, each circle measures 0 to 100, with 100 at the top, but the scale extends around to 120 or 150 if necessary in place of 20 and 50. Several pointers from the center mark the value for this month, last month, last year, etc. This method of showing index values at selected periods is quite effective, easy to read and affords correct comparisons. Another variation, instead of reproducing the entire dial, shows only a section of it, as illustrated in Figure 39. This makes it possible to use a much larger scale dial without requiring the space for a large complete circle.

FIGURE 39

DIAL CHART: INDEX OF INDUSTRIAL ACTIVITY AS OF MAY 31, 1941



Reproduced from *Buffalo Evening News*.

Linear Graphs

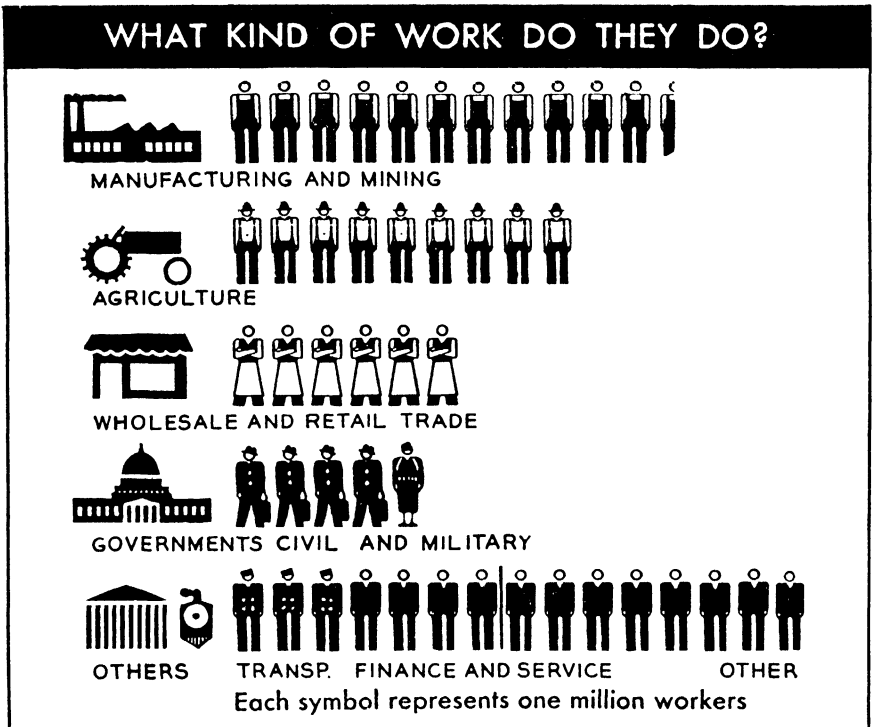
Practically any numerical value can be represented by the length of a line, and consequently bars and other adaptations of linear graphs are more widely used than any other form of statistical graph. The more complicated forms, involving time and quantitative relationships will be described in the next two chapters. This section deals only with the simpler types in which there is but one scale, which may extend either vertically or horizontally.

Pictograms.—As has already been suggested, a pictogram of the kind shown in Figure 35-D is the simplest form of linear diagram. It dates from prehistoric times, when primitive peoples drew five canoes or ten moons to represent such concepts as quantities or the passage of time. Where standardized symbols of very simple form, all the same size and evenly spaced, are shown in rows, each symbol represents a certain number of original units and several groups of them can be compared according to the lengths of the rows. The example in Figure 35-D appears to have no scale, but actually there is one, horizontally. The advantage of this graphic device as a means of distinguishing attribute classifications has already been noted. When

the method is carried to the extreme of using several variations of the symbol in each row and other symbols at the ends of the rows to indicate the different attribute classifications, as illustrated in Figure 40, the graphic requirements of clarity, simplicity, and effectiveness are likely to become lost in confusion.

Many pictogram forms that enliven the reports of governmental or other agencies do not fall within the scope of a discussion of statistical graphs; they may succeed in attracting attention but they do not convey numerical ideas. If they do aim to present statistical material the only acceptable method is the linear form, made up of identical symbols. No statistical misconception can result from such graphs, as does occur through the use of symbols or pictures of different size or those not identical in form. The rules governing the use of pictograms can be summed up as follows:

FIGURE 40
PICTOGRAM: NUMBER OF WORKERS IN BASIC FIELDS OF EMPLOYMENT, 1940



PREPARED BY PICTOGRAPH CORPORATION

1. Symbols should be self-explanatory.
2. Larger quantities are shown by a larger quantity of symbols, *not* by larger symbols.
3. Charts compare approximate quantities, *not* minute details.
4. Only comparisons should be charted, *not* isolated statements.⁶

Bar Graphs.—In Figure 35-A plain continuous bars serve the same purpose as rows of symbols. However, to most readers the bars are just as easy to understand if not more so.

Single bars: Single bars are often used to depict the separate classes of an attribute classification. They are also suitable for showing geographical classifications in which the spatial relationship of one class to another is not so important that a map is required; the three cities in Figure 35-A are an example. Such bars may also indicate values at selected periods of time that do not constitute a continuous time series. Each bar always represents a number or value for a single attribute or place or period. The bars are arranged along a base line which has no scale but which is labeled just like the stub of a table. The same principles as in tabulation determine the order of arrangement of the bars. That is, they may be in ascending or descending order of size, alphabetical, or in any other logical order.

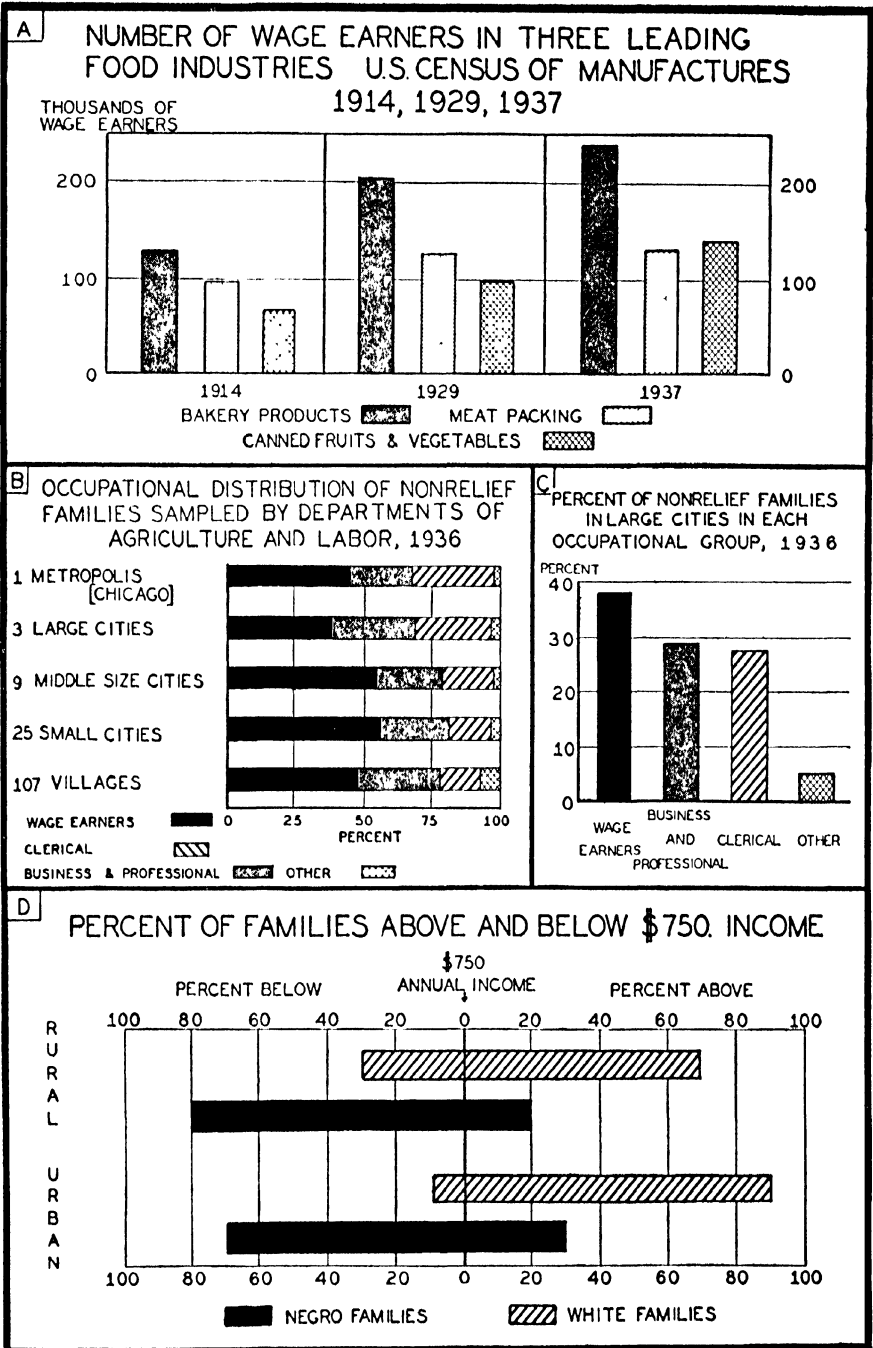
Groups of bars: As in a table, bars are often arranged in sub-classifications grouped according to whatever emphasis is desired. Solid black may be used for all the bars, whether single or in groups. When there are several subgroups, each containing the same classes of items, cross-hatching is frequently used to identify corresponding bars in each group. As in the case of circles and sectors, this hatching is merely a means of distinguishing each attribute from the others, and no order of density is prescribed except that the same order should be followed in all the groups. Figure 41-A uses this method to compare the number of wage earners in three leading food industries, at three census periods.

Divided bars: In Figure 41-B⁷ each bar is subdivided into a number of segments. This type of bar graph is used for the same purpose as the circle and sectors, to indicate parts of a whole, and the rules for cross-hatching the parts are the same as for the sectors of a circle. In planning the graph, it must first be determined which is more important, to present an accurate picture of the total values, or to afford

⁶ Rudolf Modley, *How To Use Pictorial Statistics*. New York: Harper & Bros., 1937.

⁷ Reproduced from *Survey of Current Business* (December, 1937), p. 13.

FIGURE 41
TYPES OF BAR GRAPHS



a correct cross-comparison of the proportionate distribution of parts. If the former, the original units must be shown on the scale and the bars will be of varying total lengths. If the latter, as in the illustration, the scale will be in per cents and, if all the parts are included in the graph, all bars will have the same total length, 100 per cent. When part-to-part comparisons are wanted in a single distribution the total bar may be cut into parts, each one starting from the zero base, instead of being arranged consecutively in one divided bar. This method is illustrated in Figure 41-C, which is a rearrangement of one of the divided bars in Figure 41-B. When the parts of only one total are shown, as in this case, the graph would present a better appearance if all the bars were solid black. However, if all of the bars of B were reproduced in the form of C, cross-hatching would be necessary and the entire graph would resemble the groups of bars in Figure 41-A.

Duo-directional bars: In this type of linear graph, the single scale extends in both directions from zero. It has two main uses: (1) to show percentage change among a set of comparable items some of which may have increased while others decreased, and (2) to compare values classified according to two contrasting attributes, such as male and female, Republican and Democrat. The example in Figure 41-D⁸ is a modification of this second usage and also of the types shown in 41-A, B, and C. The contrasting attributes are incomes above or below a certain standard, \$750 per year. There are two groups of bars, urban and rural, each group containing a bar for white and a bar for Negro families. Each of these four bars is actually a divided bar, that is, its total length represents 100 per cent, divided into the percentage of families having an income above \$750 per year and the percentage below \$750. The bars are aligned at this dividing point as zero instead of showing their total length measured from a common base. The method is very effective in emphasizing the particular comparison that is wanted in this case. Figure 76, page 548, is an example of a duo-directional bar diagram showing increases and decreases.

Essential features of bar diagrams: Unbroken Scale: A bar graph gives an accurate impression only when the relative lengths of the bars are correctly represented. It follows that the scale in every case must start at zero and continue as high as the highest value to be

⁸ The original form of this graph was a pictogram, in the *Consumers' Guide*, United States Department of Agriculture, September, 1938.

shown on the graph. If it starts at some point beyond zero, or if any break is made in the scale, all of the bars become shortened by *equal* instead of *proportional* amounts, and consequently a misleading picture is given of the relative total lengths of the various bars. Some statisticians consider that any labeling or figures at the far end of each bar, or on the bars, also interferes with the estimate of their lengths. However, if the figures are inserted near the zero end of the bars, there can be little objection. In the case of horizontal bars, such inserted figures appear as a column, taking the place of an accompanying table.

✓ **Shading, and Spaces between Bars:** Bars are most effective when solid black. White narrowly outlined in black is least effective since, unlike sections of a map or circle, every bar is entirely surrounded by a white background. Adjacent bars must be separated by white spaces, usually slightly narrower than the bars themselves. If they were not separated an effect of area rather than length would result, and it would be hard to estimate correctly the lengths of individual bars. Somewhat wider spaces are used to separate groups of bars, or each group may be boxed separately in a complete border, as in Figure 41-A.

Scale and Labels: In the simple types of bars thus far described there is only one scale but this scale should always be shown. There is no fixed rule as to whether the bars should extend horizontally or vertically. If they are vertical, the scale is often repeated on the right side as well as on the left, and if horizontal it may be either at the top or bottom, or both. The horizontal position is usually more convenient for graphs containing only a few bars. It is possible then to print the label of each bar at the left of the vertical base line, with the numerical value also if desired, and sufficient space can be allowed in a natural horizontal position for all labels, figures, and scale values. However, when the bars represent a time classification, even though not constituting a time series, it is customary to draw them in a vertical position as illustrated in Figure 41-A to correspond with the accepted form for time series bars, which will be explained in the next chapter.

INTRODUCTION TO TWO-DIMENSIONAL LINEAR GRAPHS

Every graph drawn on a plane surface has, of course, two dimensions. In this text, however, the simple types of linear graphs that have been discussed in the preceding section are distinguished from the

more complex types in which there are two scales of values, one extending horizontally and the other vertically.

Definitions

The term "two-dimensional"⁹ will be applied to the latter type of graphs. Whenever the data consist of series of two or more variables, a two-dimensional linear graph must be employed. Before describing the principles and methods of constructing this kind of graph, some definition of terms becomes necessary.

Variables.—This word has occasionally been used in preceding chapters, without definition. According to Day,¹⁰ "A variable is anything which exhibits differences of magnitude or number." It is used to refer to any column or row of data indicating changes in number or value of the particular unit named in the heading, e.g., in Table 15, page 149, the price of wheat is a variable. An ordered classification is also considered a variable. That is, classifications by size groups, time classifications according to regular periods or intervals, and even qualitative attributes that can be arranged in numerical order, such as age groups, are all variables. In Part A of the wheat-price table, the stub classification "time in weeks" is therefore a variable; the classifications in Parts B and C are not variables, however, since neither "grades of wheat" nor "market" depend upon differences in magnitude or number.

Statistical Series; Dependent and Independent Variables.—When two such variables are shown in relation to one another the resulting table of data is a statistical series. Thus there may be series in time or series according to attribute.¹¹

Variables in a series are further defined as "dependent" and "independent," the unit usually being the dependent variable and the classification the independent variable; e.g., the price of wheat is the dependent variable and time in weeks is the independent variable. Under some conditions both variables could be considered independent, in which case either one could be classified in terms of the other. For example, instead of quoting prices according to the time classification,

⁹ Not to be confused with "duo-directional" in which a single scale extends both positively and negatively from zero; nor with "double or multiple scale," a term to be introduced later referring to two or more scales of units both measured vertically.

¹⁰ Edmund E. Day, *Statistical Analysis* (New York, 1925): The Macmillan Co., p. 10.

¹¹ Series in space are also suggested by Day (*ibid.*, p. 45), but this possibility seems not wholly consistent with his definition of a variable, since no geographic classification "exhibits differences in magnitude or number."

weeks, a frequency distribution of the same data might use price range as the stub classification, the unit being the number of weeks in which each price was quoted:

PRICE RANGE	NUMBER OF WEEKS
\$.60-\$.649	3
.65- .699	10
.70- .749	22
etc.	

In the discussion of classification in chapter VIII it was stated that from the point of view of tabular arrangement no distinction need be made between series and other kinds of classification. In graphic development, however, the treatment for a series of two variables, whose relations to one another are of primary importance, is quite different from the methods that have been described in this chapter for a single variable classified non-numerically.

Two-Dimensional Scales.—In order to measure graphically the values of the two variables in a series, two scales are required on two axes that intersect at right angles. The vertical axis corresponds to the single scale described in the preceding section on linear graphs and is used to measure the number of units in the dependent variable, that is, the numerical values of the data themselves. In other words, it represents the figures that appear in the rows and columns of a table. The horizontal scale in similar fashion measures the independent variable, that is, the numerical values of the classification.¹²

Ordinarily such a graph uses only one quarter of the field covered by the intersecting axes, that is, the "positive" field above and to the right of the point of intersection. If negative values are necessary on the vertical scale, the field below the intersection must also be shown, and similarly the left-hand field for negative values on the horizontal scale. For use in formulas and diagrams, the dependent variable is referred to as *Y* and the independent variable as *X*. The *Y* variable is called a function of the *X* variable. The student will do well to familiarize himself with all these uses of the term "variable" because they become increasingly important in advanced statistical analysis.

Types of Statistical Series and Their Graphs

When the classification indicates quantitative attributes, the units in the dependent variable are called "frequencies." Graphic methods for

¹² Certain exceptions to this rule will be found in some special purpose graphs, such as price curves in the field of economics.

illustrating analysis of this kind of series will be described in chapter XV. Series in which two quantitative attribute classifications are related to one another lead to analysis by correlation, the subject of chapter XXVII. The principles of time series and the methods for representing them graphically are discussed in the first half of chapter XIV.

In the majority of texts it is customary to reserve the use of the word "series" for time series, and to refer to other kinds of series as "groups." Consequently that practice will be adhered to in subsequent chapters of this book. The general term "series" was introduced in this section to point out a basic similarity in all data involving functional relationships between two variables. A two-dimensional scale is necessary for graphic representation of any such relationships.

PROBLEMS

1. What is the principal advantage of a graph as contrasted with a table for presenting information?
2. State briefly the relations shown in each part of Figure 35, page 299.
3. Find in print a graph of one of the types presented in Figure 35. Analyze on the basis of our text the steps followed by the author of the graph in its preparation.
4. Select from any issue of the *Statistical Abstract* tables which should be represented by each of the four types of statistical maps described in the text. Explain why each set of data could be represented best by the type of map you have selected.
5. a) Present the information given below in the form of circles and sectors.
b) Discuss the difference in use of fuels by iron and steel, and by all industries.

COST OF INDUSTRIAL CONSUMPTION OF PURCHASED ELECTRICITY AND OTHER TYPES OF FUEL, FOR ALL INDUSTRIES, AND FOR IRON AND STEEL, 1929*

TYPE OF FUEL	COST (MILLIONS OF DOLLARS)	
	ALL INDUSTRIES	IRON AND STEEL
Total	1,973.9	463.1
Purchased electricity	719.5	128.2
Bituminous coal	754.5	87.2
Anthracite coal	43.6	2.5
Coke	243.7	198.2
Fuel oils	212.6	47.0

* 1930 Census of Manufactures, Vol. 1, p. 161.

6. Given the following information concerning the age distribution of all persons over 10 years of age in the United States and those gainfully employed in each group:

	MILLIONS		
	1900	1920	1930
Total population 10 years of age and over.	57.9	82.7	98.7
Total gainfully employed.....	29.1	41.6	48.8
Total population 10-15 years of age....	9.6	12.5	14.3
Total gainfully employed.....	1.7	1.1	.7
Total population 16-44 years of age.....	34.7	48.1	56.3
Total gainfully employed.....	20.2	28.9	33.5
Total population 45-64 years of age.....	10.4	17.0	21.4
Total gainfully employed.....	5.8	9.9	12.4
Total population 65 years of age and over..	3.1	4.9	6.6
Total gainfully employed.....	1.2	1.7	2.2

- a) Study the changes in the age composition of the employed population during the 30-year period, and draw a graph to illustrate your conclusions.
- b) Study the changes in the percentage of each group gainfully occupied, and draw a graph to illustrate these changes.
- c) Write an interpretation of the data illustrated by your graphs.
7. a) Present the following information graphically.
- b) Discuss the nature of changes in this business during the ten-year period.

SALES IN A COUNTRY GENERAL STORE ACCORDING TO
TYPE OF GOODS, 1930 AND 1940

TYPE OF GOODS	1930	1940
Groceries	\$19,650	\$21,410
Meats	400	975
Shoes	1,125	630
Rubber footwear	760	925
Dry goods	1,650	310
Notions	1,850	1,025
Hardware	925	1,070
Drugs	425	115

REFERENCES
(See page 348)

CHAPTER XIV

GRAPHS—*Continued*

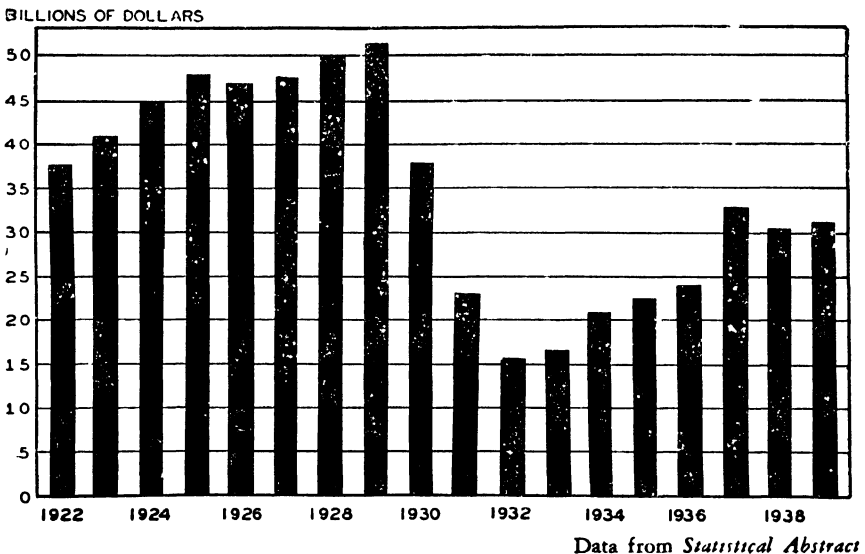
TIME SERIES GRAPHS

THE passage of time is most naturally pictured by a moving point whose apparent course may be traced from left to right. This basic idea is utilized by all types of time series graphs.

Bar Graphs

A time series may be represented by a row of bars, the height of each bar representing a single value, exactly as in the one-dimensional graphs. However, the order of arrangement no longer depends on judgment or arbitrary choice; the bars must stand at evenly spaced intervals along the base scale, the units of which represent successive equal periods of time. The fluctuations of the dependent variable may be followed through the path marked by the tops of the bars. This is illustrated in Figure 42, which shows the changes in value of United States exports annually, 1922–39.

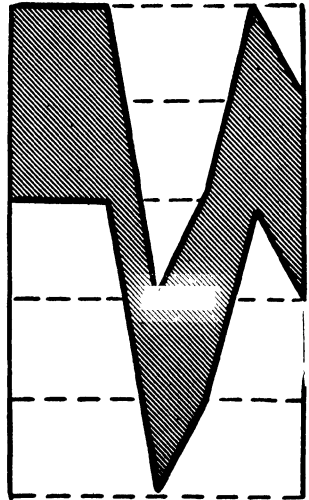
FIGURE 42
BAR GRAPH OF TIME SERIES
VALUE OF UNITED STATES EXPORTS, 1922–39



Band, Strata, or Surface Graphs

Bars that represent a time series may be divided into several components. A long row of divided bars gives an effect of wavy horizontal bands or strata. To serve the same purpose, these several component parts of the variable are frequently shown as a continuous "surface" instead of in separate bars. The strata or bands in the surfaces are cross-hatched the same as divided bars, and show by contrast the changing proportions of the parts of a total over a continuous period of time.

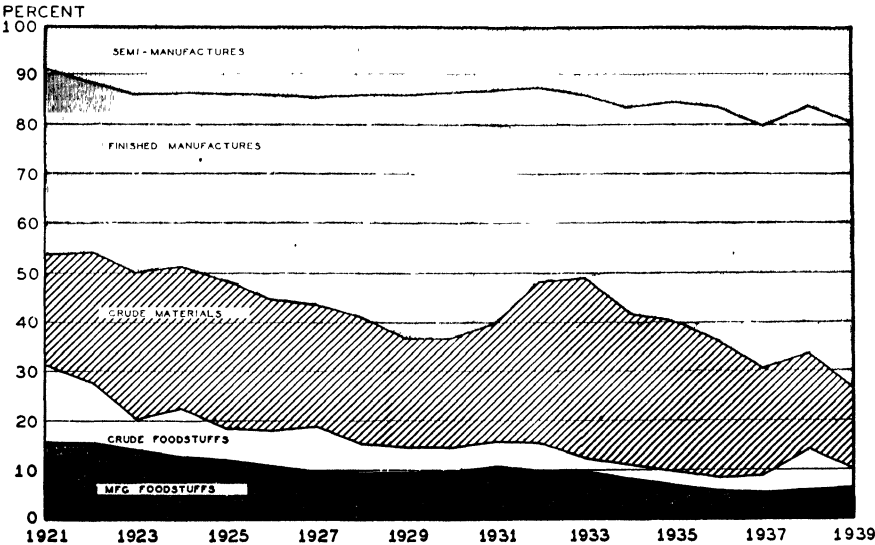
This type of chart may be designed to show percentage distributions, in which case the graph consists of a rectangle completely filled in with bands of fluctuating width. If the scale represents actual values instead of per cents, the upper boundary of the surface will be irregular, representing the actual total at each period. The two types are illustrated in Figure 43-A and 43-B, both of which show the same total data as Figure 42, divided into component parts. Whether per cents or actual values are shown, there is some danger that the bands may take on a distorted appearance due to sudden extreme fluctuations in some of the parts. The width of the band must be estimated by the *vertical* distance between its boundaries at each point. In the illustration at the right the width is actually the same throughout the period, but due to various angles of change in its lower boundary, it is pulled out of shape. If the data do not contain too many sudden changes, this distortion may be reduced to a minimum by charting at the bottom of the graph the narrowest band and others having but slight fluctuation, so that succeeding bands will have a lower boundary that is fairly level. The upper one, that which fluctuates the most, will then not affect the shapes of the other layers.



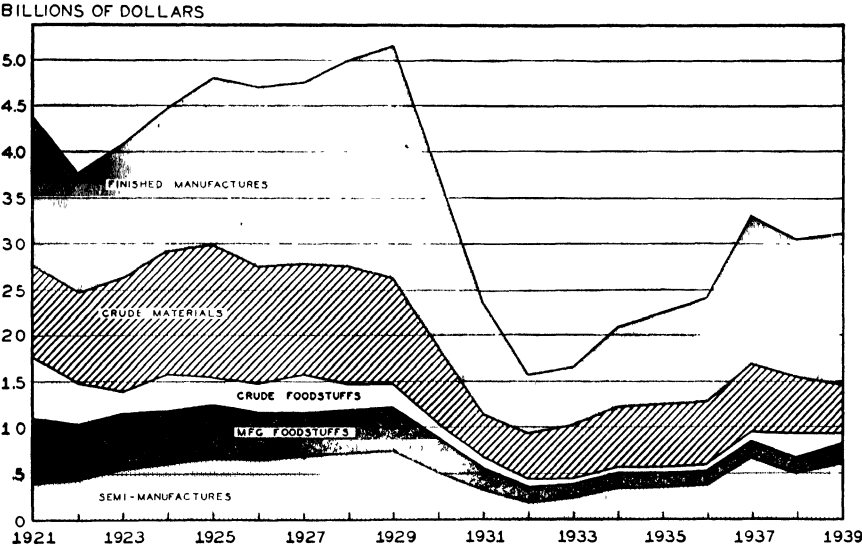
The smoother strata can be located near the top and bottom in a 100 per cent graph. The jagged edges of the two most irregular strata will then fit together near the center, their widths being measured from the straighter edge of each. In some cases, however, a required sequence of the component parts based on other considerations of the data will determine the order of the bands.

FIGURE 43
BAND GRAPHS OF TIME SERIES
VALUE OF UNITED STATES EXPORTS ACCORDING TO MAJOR ECONOMIC CLASSES,
1921-39

A. CUMULATIVE PER CENTS



B. CUMULATIVE DOLLAR VALUES



Data from Statistical Abstract

A well-planned band, strata, or surface chart is a valuable means for showing the changing relations of the component parts to the whole in a time series. On the other hand, if the primary purpose is to place emphasis on time comparisons between individual parts, or between separate totals, a line graph is preferable.

Line Graphs of Time Series

Graphs of this type are unquestionably more widely used than any other in every phase of government and business statistics. They are constructed in exactly the same way as simple bar graphs of time series, except that, instead of drawing a bar for each time period, only the point at the upper end of each bar is plotted.¹ The successive points are then connected by straight lines whose combined length may have a more or less jagged appearance, depending on the irregularity of the data. Regardless of its degree of smoothness, this continuous line is called a "curve." The chief advantage in using curves instead of bars is that several curves may be shown for comparison on the same graph more easily than several sets of bars.

Curves of time series may serve either of two purposes: to show actual amounts of change or to show relative changes. Whichever function is more important will determine whether the vertical scale should be arithmetic or logarithmic.

Arithmetic Scale.—An arithmetic scale in which equal spaces on the scale stand for equal amounts of the unit is familiar to everyone. It should be used for the vertical scale whenever a comparison is wanted between actual amounts of the unit either for a single series at different time periods or for several series at corresponding periods.

Methods of comparing several series: A problem arises, however, in comparing two or more series that are recorded either in different units, or in the same unit at levels so far apart that it is difficult to use the same scale effectively for both. The purpose of such a graph must be carefully considered before choosing one of the alternative graphic methods for dealing with this situation. Figure 44, A, B, and C, shows three ways of handling the same data on arithmetic scales.

¹ In mathematical language, the positions assigned to values of the independent variable along the horizontal axis are called "abscissas" and the values of the dependent variable assigned along the vertical axis are called "ordinates." The plotting of any value of the dependent variable consists in (1) determining the position of the independent variable on the horizontal scale (abscissa) and the value of the data on the vertical scale (ordinate); (2) locating the point of intersection of a vertical at the abscissa position and a horizontal at the ordinate value.

1. **Single Unit Scale:** If the primary purpose is to compare absolute amounts at each period, a single² unbroken arithmetic scale is best even though it minimizes the fluctuations of the series having the smaller values. In this case, if the purpose is to show that coal is still the major source of power, as compared with oil, Figure 44-A is the one to use.

2. **Index Numbers:** If a comparison of relative changes of the variables will serve the purpose, and particularly when two or more series do not have a common unit, they may be reduced to indexes, using a corresponding base³ period in each case. The 100 per cent line and the entire per cent scale will be common to the several series. Figure 44-C shows the relatively greater percentage variation of oil as a source of power, 1906-10 being chosen as the base period.

3. **Scale Equation:** If it is essential to depict on the graph the actual rather than the relative values, and at the same time to show on equal terms the degree of fluctuation in each series, some form of scale equation must be utilized. Several methods are commonly used but the only one that is justified statistically⁴ is the equation of the several series of values so that their respective averages for the period will approximately coincide on the graph. The method is as follows: (a) Find the arithmetic average of each series, in its own unit, for the entire period. (b) Find the ratio between the two averages. In this example, the averages were 8.7 quadrillion B.T.U. for coal consumption and 1.62 quadrillion B.T.U. for oil, or roughly one unit for oil to five for coal. (c) Separate vertical scales are drawn on either side⁵ of the graph, both starting at zero. In Figure 44-B, coal is on the left and oil on the right. (d) The same space that represents five units on the left (coal) represents one unit on the right (oil), and the approximate averages of the two coincide. (e) In arranging the equated scales, they should not be condensed so much that fluctuations are underemphasized, but the maximum values of both sets of data must be provided for.

² This method, of course, is available only when the two series are in the same unit or can be reduced to a common unit. Otherwise the comparison of absolute amounts on a single scale is out of the question.

³ For the choice of base, and interpretation of the comparison, refer to the discussion of index numbers in chapter XIX, pages 483-86 and page 498.

⁴ The suggestion has been made by some statisticians that the scales be equated on the basis of dispersion. However, the entire purpose of scale equation is to compare the degree of fluctuation between two series of data, and if this dispersion is equalized there appears to be no useful comparison shown by the graph.

⁵ If three series must be equated it is necessary to have three separate scales, each properly labeled. Customarily two of the scales are placed at the left and one at the right, but occasionally all three will be found at the left. With more than three series the graph becomes too involved to read easily.

FIGURE 44

LINE GRAPHS OF TIME SERIES

SUPPLY OF POWER FROM COAL AND DOMESTIC OIL
ANNUAL AVERAGES FOR FIVE-YEAR PERIODS, 1871-1935
(All figures represent equivalent British Thermal Units, in quadrillions)

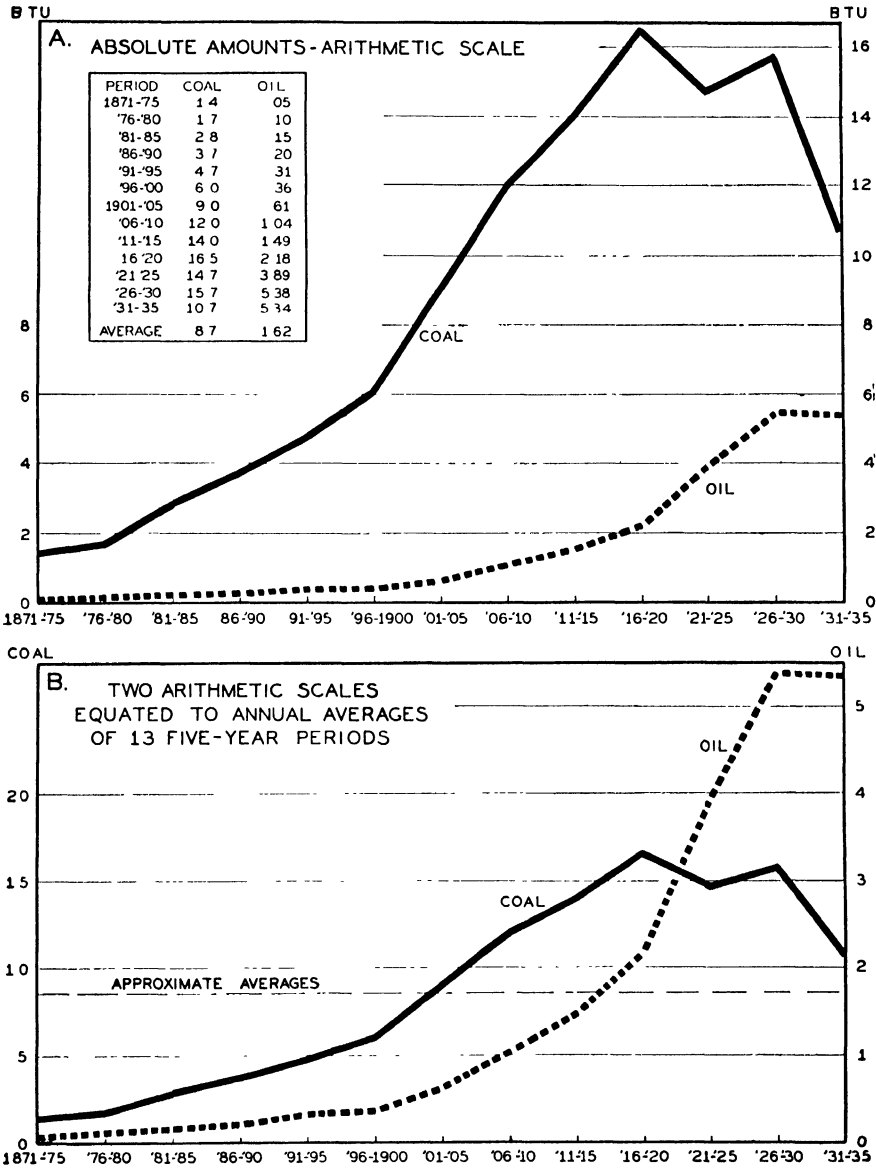
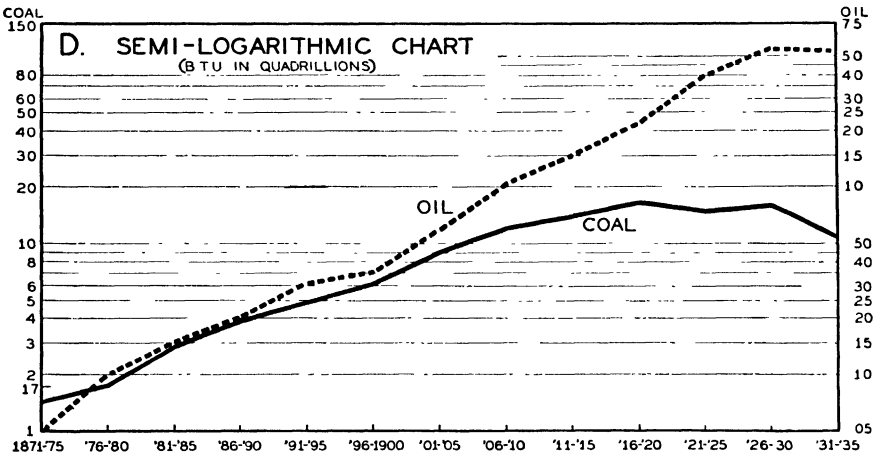
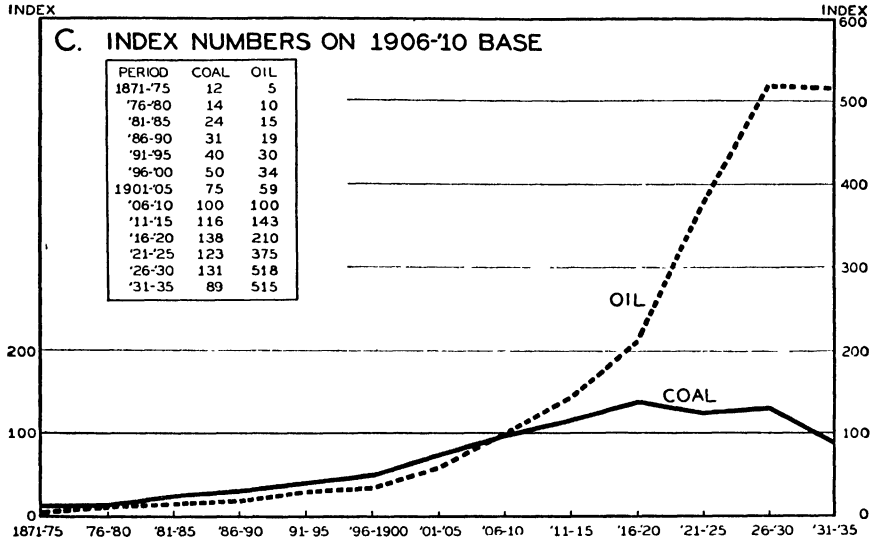


FIGURE 44 (continued)
LINE GRAPHS OF TIME SERIES



Data from *Statistical Abstract*

As a result of this method, each variable is given equal emphasis in terms of fluctuations from its own average, and they may be compared accordingly. For example, Figure 44-B shows that after 1916-20 oil-produced power expanded sharply above its average level, while in the same period coal-produced power fluctuated mildly with a slight tendency to decline toward its average level. The importance of the

increase in oil-produced power during this period is completely concealed in Figure 44-A. Specifically the increases in coal-produced and oil-produced power appear to be equally important in 44-A but 44-B shows the actual relation in the growth of use of the two fuels.⁶

It should be noted that if index numbers are computed on an average of the period as a base, the shapes and relative locations of their curves will be exactly the same as by this more cumbersome method of scale equation. The only advantage of scale equation over index numbers is that actual amounts instead of percentages can be read from the graph.⁷

4. Need for Logarithmic Scale: A comparison of relative rates of change, or relative increases or decreases from one period to another is likely to be more important than any of the three purposes named above. Such comparisons cannot be shown satisfactorily on an arithmetic scale, even by means of index numbers. Changes in the latter must always be studied with reference to a certain base period and cannot be shown equally between any two periods or at various levels on the scale. This is because equal amounts, or spaces, on an arithmetic scale represent constantly decreasing percentage changes as the values of the scale increase. For example, in Figure 44-C the index of coal consumption increased from 14 to 24 from 1876-80 to 1881-85, a relative increase of over 70 per cent, but the index rose only ten spaces on the scale; on the other hand, from 1921-25 to 1926-30 the oil consumption index rose from 375 to 518, an almost identical percentage increase (74 per cent), but 143 spaces on the scale were required to show it. In order to give an accurate visual conception of these relative rates of change, the arithmetic scale must be abandoned in favor of the logarithmic. This fourth method is shown for the oil-coal data in Figure 44-D and will be explained later.

Breaks in time series scales: Vertical Scale: Strict accuracy requires that there should be no break in the arithmetic scale of a time series line graph any more than in the case of bars. However, it is often just as important to study the fluctuations in the variables as to compare the actual total values between the points of the curves. The lowest value of any of the variables may amount to several million

⁶ Students will be able to make this type of interpretation in greater detail after studying measures of dispersion in chapter XVIII.

⁷ Note Figure 78, page 555, in which two scales are equated to the values at the first period instead of to the respective averages. In this case the purpose of the graph is to show how the two series diverge.

units, and a scale covering a complete range between zero and the highest value to be graphed would become so small that a change of even several thousand units would cause no perceptible movement in the curve. Common practice, therefore, permits a "break" in the vertical scale below the lowest point needed for any of the points plotted. Zero is indicated as the base, then a double jagged line is drawn (using finer lines than the curves on the graph) and the scale may be resumed above the break at any value required. The break represents an actual tear in the paper, hence the vertical scale line and all grid lines are left blank between its boundaries. (See Figures 73-A, 74 and 75, pages 540, 543, and 546, respectively.)

Likewise when the values of the series are in the form of index numbers, if the vertical scale is incomplete, a more careful study becomes necessary in order to estimate correctly the percentage of fluctuation. However, in order to enlarge the per cent scale, zero is frequently omitted altogether, the scale being extended below 100 only as far as the data require. In any case, the 100 per cent or normal line should be emphasized, since it is just as important a standard as zero.

Horizontal Scale: So much emphasis has been placed on the fact that the regular intervals of a time series graph accurately depict the even progress of time that any suggestion of tampering with this scale may well be questioned. There is no abrogation of the principle that equal spaces on any arithmetic scale should always represent equal values. However, just as under some circumstances a break may be permitted in the vertical scale, there are also situations which justify changes rather than breaks in the time scale.

In many business charts the main interest is in the presentation of current monthly or weekly data. Comparison with the past is a matter of only secondary importance, and it has therefore become customary to enlarge the scale for the current year and to contract the scale for previous years. If space is limited, and it is desirable to show some information regarding changes in the past for as long a period as possible, there are several methods for representing the earlier years in condensed form. Some possible alternatives are: (1) to show on a small scale monthly changes for several previous years; (2) to represent only the annual averages by a single point in each year of the earlier period; (3) to use a series of vertical bars each of which indicates the range of fluctuation within a single year. Any of these meth-

ods gives a more complete and continuous story than can be gained from a chart in which there are either no data at all regarding previous periods, or else there is a complete break of several years with no indication of what happened in between. For the business man who has become accustomed to these forms there is no misrepresentation of facts. He has in compact form just the information he wants, and is well aware that the current year is a sort of "slow motion picture" in comparison with the previous scale. However, these procedures are not recommended for the student as methods for general interpretation. When they are used, the chart should be divided vertically into several segments separated by narrow white spaces to indicate the points at which the base scale has been changed. Figure 71, page 531, illustrates two changes in the time scale. The main interest is in the weekly data shown on a large scale for 1940 and 1941; the preceding four years of monthly data are shown on a smaller scale; and for the earlier years the bars show the annual range and the average for each year.

Logarithmic Scale.—The logarithmic⁸ or ratio scale is widely used as a graphic device because it permits equal spaces to represent equal percentage changes at any point on the vertical scale of a time series. The space between 100 and 200 is the same as between 50 and 100, 20 and 40, 6 and 12, 1.5 and 3, .005 and .01, and so on.

Explanation of principle: Figure 45 illustrates the method by which the spacing on the logarithmic scale is determined. On the left (A) is an ordinary arithmetic scale, from 0 to 2; in the center (B) is a column of logarithms, whose characteristics range from 0 to 2, marked off at points measured according to the arithmetic scale (A); on the right (C) is a column of natural numbers which are the anti-logs of the logarithms opposite them on scale (B). These natural numbers in (C) are therefore spaced according to what is known as the logarithmic or ratio scale.

The advantage of using scale (C) is based on the rule for multiplication by means of logarithms: when two numbers are to be multiplied together, their logarithms can be added and the sum will be the logarithm of the product of the two numbers. In Figure 45 the space marked $a-c$ on the center scale stands for the logarithm of 6; $a-b$ stands for the logarithm of 2. Hence if we wish to multiply 6 by 2

⁸ See Appendix C for explanation of logarithms, rules for their use, and table of logarithms.

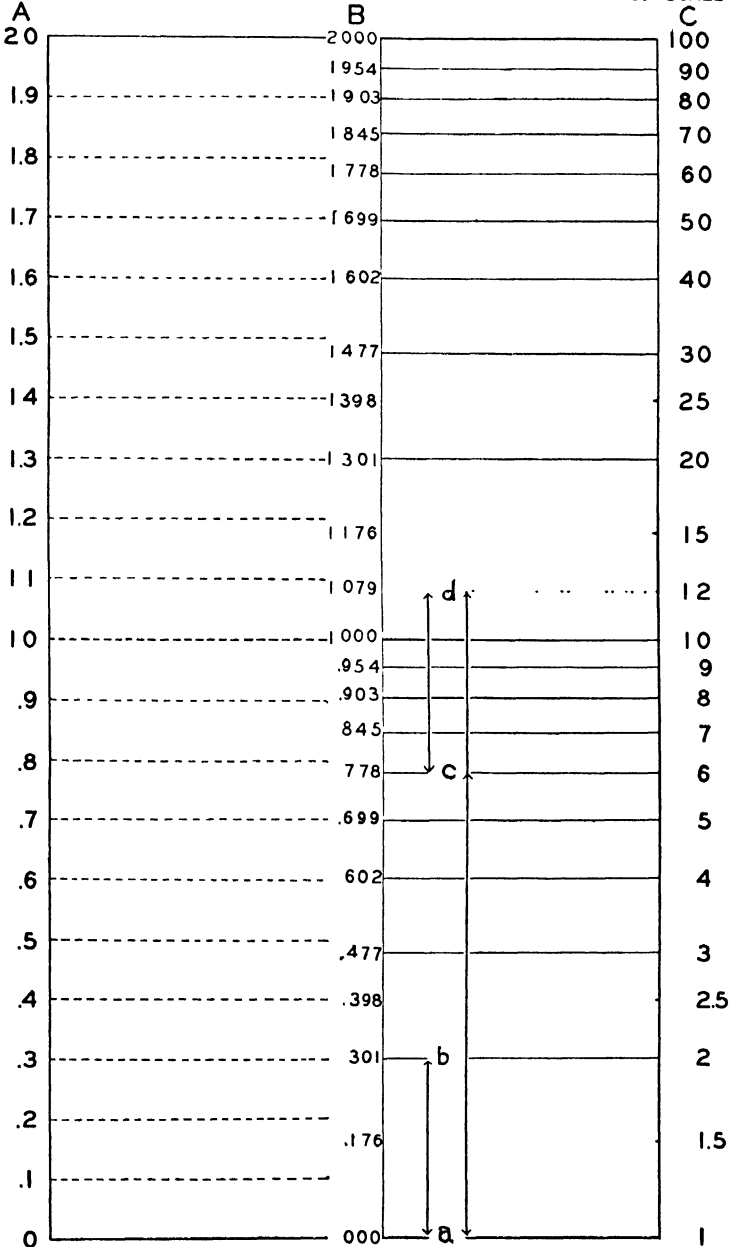
FIGURE 45

CONSTRUCTION OF THE RATIO SCALE

ARITHMETIC SCALE USED
IN SPACING LOGARITHMS
OF SCALE B

LOGARITHMS
MEASURED
BY SCALE A

NATURAL NUMBERS
ANTILOGS
OF SCALE B



(that is, to double it or increase by 100 per cent) we can add a space $c-d$, equal to $a-b$, to the space $a-c$, and we should arrive at $a-d$, the logarithm of 12. This is precisely what does occur, as can be verified by measuring with a ruler. Likewise the space from log 4 to log 8 can be measured and proved equal to the same space $a-b$, as is also log 75 to log 150, log 300 to log 600, etc. In other words, adding the space $a-b$, representing log 2, to any other value at any point on the scale (C) will multiply that value by 2, or increase it by 100 per cent.

Now take the space on scale (C) measured by log 3. Added to itself, we reach log 9; added to log 10 we reach log 30; etc. In each case the original value of the anti-log is multiplied by 3 or increased by 200 per cent.

The space measured on scale (C) by 1 to 10, or 10 to 100, is called a cycle. Every point in a cycle is ten times the value of the corresponding point in the cycle just below it, or represents 900 per cent increase.

Percentages of decrease follow the logarithmic rule of division: the quotient is the anti-log of the difference between the logarithms of two numbers. Just as in the case of any percentage change computation, per cents of increase or decrease between two given points have different bases and must be read differently. That is, log 50 plus log 2 = log 100, an increase of 100 per cent. But log 2 subtracted from log 100 = log 50; 50 is $\frac{1}{2}$ of 100, a decrease of 50 per cent. Similarly log 3 subtracted from log 90 = log 30; 30 is $\frac{1}{3}$ of 90, a decrease of 66 $\frac{2}{3}$ per cent. And log 50 subtracted from log 200 = log 4; 4 is $\frac{1}{50}$ of 200, a decrease of 98 per cent.

In the portion of logarithmic scale illustrated in Figure 45, only two cycles are shown, from 1 to 10 and 10 to 100. This scale can be extended upward, of course, to 1,000, 10,000, etc., and it may also be extended downward indefinitely to .1, .01, .001, .0001, but never can reach zero.⁹ There is therefore no zero base on this scale, nor any other fixed point from which heights are measured. Hence only the portion of the scale that is used in plotting need be shown in the graph of a given series. Relative changes are measured by the distance between any two points on the vertical scale. Any alteration in the

⁹ Consequently the logarithmic scale cannot be used for a series that includes zero or negative values

slope of a curve, therefore, indicates a changing relative rate of change¹⁰ in the data. A curve that follows a straight line upward is increasing at a constant relative rate. The possibilities of changing relative rates of increase in a curve, and corresponding relative rates of decrease, are illustrated in Figure 46.

- a) If it is convex upward, it is increasing at an increasing relative rate.
- b) If it is concave upward, it is increasing at a decreasing relative rate.
- c) If it is concave downward, it is decreasing at an increasing relative rate.
- d) If it is convex downward, it is decreasing at a decreasing relative rate.

Because there is no fixed base line, scale equation between differing units presents no great problem. One unit can be shown in tens on one side of the scale, and another unit in thousands on the other side. The scale values can be adjusted at will in order to bring the curves to the relative positions that afford the most effective comparison of their slopes at various periods, *provided the original ratio relationship set up by the logarithmic scale is not tampered with*. This means that every value may be *multiplied*¹¹ by the same number throughout, changing the cycles, for example, from 1-10-100 to 3-30-300, or 4-40-400. Each cycle is still ten times the value of the cycle below it, and the intervening spaces also keep their original ratio values. It is sometimes convenient to make this adjustment by multiplying in order to bring the curves closer together or to bring all the values within one less cycle. For example, the series 8, 20, 36, 47, 80, 200 would require the use of three cycles 1-10-100-1,000. But if the scale is multiplied by any factor from 2 to 8, the series will fall within two cycles, 2-20-200 or 8-80-800.

Example and interpretation: The advantage of the semi-logarithmic time series graph, that is, one in which the time scale is arithmetic and the vertical scale logarithmic, can be illustrated by a study of

¹⁰ When we speak of rates of change on a logarithmic scale, those rates are expressed in per cents. They are really relative rates of change and the fact that they are expressed as per cents is usually taken to imply relative rates without employing the cumbersome terminology. However, we shall use "relative rates of change" in this text to keep the student constantly reminded that the rates of change are expressed in per cents.

¹¹ Note that if the original values are *moved up* on the scale, as moving 1 to 2, 2 to 3, etc., or if an equal amount is *added* to each original value, the true relationship of the logarithmic values will be entirely distorted

FIGURE 46

CURVES SHOWING CHANGING RELATIVE RATES ON A RATIO SCALE

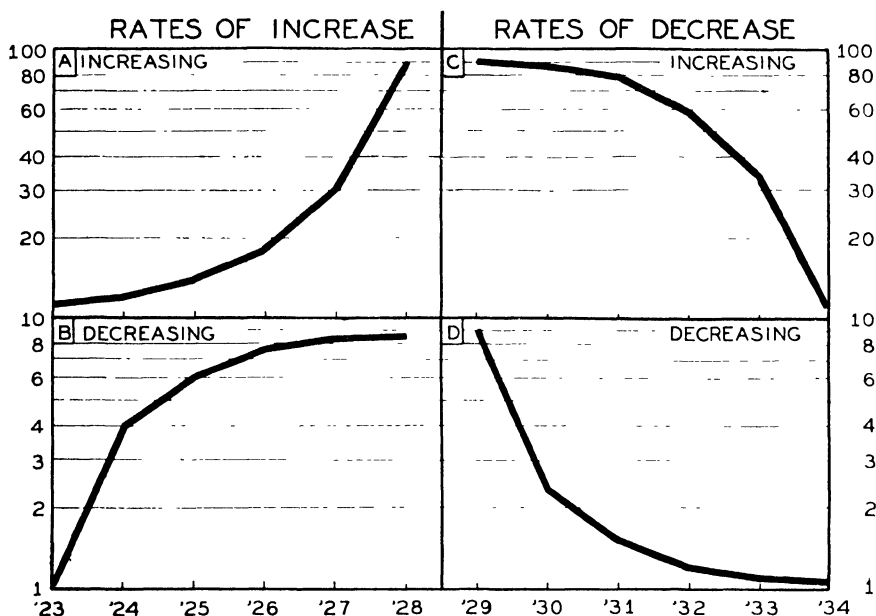


Figure 44-D. After looking at the three preceding graphs of these data, drawn on three different arithmetic scales, one scarcely knows which fuel has increased in use the more rapidly. The rates of change could be computed from A or B but certainly they are not readily apparent on either graph. From Figure 44-C which shows the index numbers, the percentages of change can be compared as related to a certain base period. This indicates that the use of oil has increased faster than that of coal, but if some earlier period had been taken as the base the relative increase in the use of oil would have been greatly exaggerated. With the logarithmic scale, however, the difference in the slopes of the two curves is apparent at a glance.

From the first to the second five-year average we know that oil-produced power increased at a greater rate than coal-produced power because its curve slants upward more sharply. For the second five-year interval the use of coal increased a little faster than that of oil, but during the next period the two curves are practically parallel, hence the two were increasing at an equal relative rate. From this time on until 1926-30 the use of oil increased more rapidly during every period

except 1891-95 to 1896-1900. During the final period the consumption of both fuels declined, but coal more sharply than oil.

Coal-produced power increased at an almost constant relative rate from 1881-85 to 1906-10, then it leveled off slightly for two periods, and finally turned downward. Its subsequent course has been a decrease at an increasing relative rate except for a minor recovery in 1926-30. Oil-produced power likewise increased at a constant relative rate from 1896-1900 to 1906-10; the line then leveled off, indicating a smaller relative rate of increase for the next two periods, after which it turned up and resumed its previous course for one more five-year period. This periodic rise in the curve can be measured on the scale and found to be equivalent to the distance between 1.0 and 1.7, or a 70 per cent increase. Computation either from the original data, Figure 44-A, or from the index numbers, 44-C, will prove this to be approximately correct for the three periods, 1896-1900 to 1901-5, 1901-5 to 1906-10, and 1916-20 to 1921-25.

It scarcely needs to be added that since the slopes of the curves are so significant in this type of graph, neither of the scales can be tampered with in any way. Any omission of intervals or change in the time scale would entirely distort the slopes of the curves.

Methods of making a ratio scale: Semi-logarithmic graph paper can be purchased in any needed number of cycles and with almost any arrangement of the base scale in time intervals. However, in order to use the ratio scale confidently, the student should understand how it is made and should be able to make his own scale if necessary. It can most easily be marked off with a slide rule if one is available. If a table of logarithms is at hand the simplest procedure is not to draw a complete ratio scale, but to plot the logarithms of each point on an arithmetic scale, just as scale (B) was plotted according to scale (A) in Figure 45. The plotted points will be equivalent to the anti-logs (natural numbers) plotted on a ratio scale, just as scale (C) was equivalent to scale (B). If neither of these aids is to be had, correct proportions may be obtained by plotting a geometric series¹² at evenly spaced intervals on the vertical scale using any starting point (t) and any common multiplier (r). Thus if $t_1 = 1$ and $r = 2$, a scale of 1, 2, 4, 8, 16, 32, 64, 128, 256, etc., could be used for

¹² In a geometric series the ratio of any term to the preceding term is constant. If t_1 denotes the first term, n the number of terms and r the common ratio, the series may be written,

$$t_1 + t_1r + t_1r^2 + t_1r^3 + \dots + t_1r^{n-1}$$

plotting. Although the scale is accurate, the plotted curve may be somewhat approximate because it is hard to determine exact values on such a scale.

PLANNING GRAPHS FOR GENERAL EFFECT

After having selected the kind of graph he intends to use to illustrate his point, the statistician is ready to block out his actual plan for drawing. His degree of success at this point will be in direct proportion to his ability to combine artistic principles and technical skill with statistical acumen.

Artistic Considerations

Fortunately, no artistic genius is required in order to create an artistically effective graph. It is necessary only to understand the simplest rudiments that serve as guides to practically every form of artistic expression—size, proportion, balance, and contrast.

Size.—The size will depend primarily on how the graph is to be used: is it to be published, or used for lecture purposes? A wall chart must be large and clear enough to be seen from any point in the room or auditorium. There is no use in preparing such a chart if it is too small. The lighting conditions under which it is to be shown must also be taken into account.

If the graph is to be printed the size of the page will determine its final dimensions, but the original may be drawn from $1\frac{1}{2}$ to 3 or 4 times larger. Less meticulous care will be needed to draw it on a scale larger than its final form since small imperfections will disappear in photographic reduction.

The amount of detail included is also a factor in determining the size of a graph. If only a single important relationship or a general condition is to be emphasized, one-half or even one-quarter of a page will suffice; whereas a more important or comprehensive graph will require a full page. If the graph includes a great deal of complex information, a variety of different kinds of lines, and detailed scales and legends, it may be necessary to use a folded insert even larger than the page of the book.

Students who use prepared graph paper of standard size should remember that it is not necessary always to make a graph that will cover the entire sheet. Each graph can be made of suitable size and proportion by inclosing a part of the page within a border.

Proportion.—The exact relation between the length and width of a graph is determined to some extent by the data that are being presented, but in general there is a range within which a pleasing effect may be attained. If a graph is too long and narrow, either horizontally or vertically, it has an awkward, stretched-out appearance. Square graphs present a monotonous appearance and do not fit the page conveniently. The proportions will be within a pleasing range if the length is somewhere between $1\frac{1}{2}$ and $1\frac{3}{4}$ times the width.¹³ Probably the most convenient standard to use in preparing material for publication is known as "root two" or the ratio of 1.414 to 1. The long side is equal to the diagonal of a square drawn on the short side, and consequently if the rectangle is divided in half the resulting rectangles have the same proportions as the original, i.e., 1 to .707. A graph that is drawn with these proportions may occupy a whole page turned the long way, or reduced to half size will fit across half of the same space in normal position.

Balance.—The term "balance" as applied to a graph has the same meaning as in any other kind of picture. It is a term borrowed from physics to indicate that there is an approximately equal stress on either side of a central point. The statistician is not at liberty to select his data so that, for example, the peaks and troughs of his curves will balance artistically. He must therefore depend upon his auxiliary material if necessary to offset the appearance of an unbalanced set of data. To a certain degree he can enlarge one scale and reduce the other in order to alter the shape of his curve, although discretion must be used to avoid an exaggerated effect of fluctuation. If he has a set of bars that nearly fill the entire space, he will make them slender enough so that they will not bulk too large and heavy. If in spite of every effort a good fourth or more of the surface remains blank, he might print his title or key in that section, or insert a small table of the data. (See Figure 44-A and C.) Note, however, that *it is never permissible to insert printed material between any significant portion of the graph and its accompanying scale*. Instead of using a key, if legends are printed close to each curve, it is usually possible to distribute them in clear spaces on the graph instead of bunching them all at the top or at one side. The addition of a border is a great aid in tying all parts of a graph together into a well-balanced whole, and is partic-

¹³ "Length" refers to the longer dimension which is most frequently the horizontal measurement; "width" is the shorter dimension, usually the height.

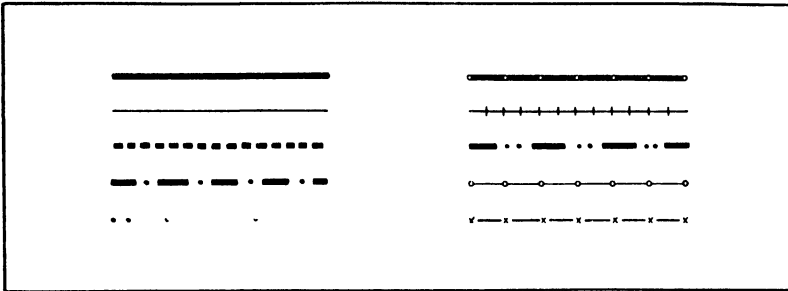
ularly necessary for maps, circles, and simple bars. (See Figures 35 and 36, pages 299–305.) Many graphs will be found in print that have no borders except for the page margins. These graphs are usually of the two-dimensional type, such as Figures 42 and 43, in which the limits of the grid itself bounded by the horizontal and vertical axes with the title of the graph printed above practically take the place of a border in marking out the definite space occupied by the graph.

Contrast.—Boldness is the secret of effective graphic presentation. Since the sharpest contrast can be achieved by the use of black and white, this combination is most commonly used in statistical graphs. Other reasons for preferring black and white are: (a) Graphs in color are much more expensive to reproduce, and some colors cannot be used in ordinary photostating. (b) Colors cannot be arranged in varying degrees of intensity as unmistakably as can the standard types of black and white cross-hatching. (c) All readers do not evaluate colors in the same way, and some may even be color blind.

Appropriate types of cross-hatching for various purposes have already been discussed and illustrated in chapter XIII by the circle in Figure 35-B, the map in Figure 37, the bars in Figure 41, and in this chapter by the strata charts, Figure 43. To sum up the general rule again: when cross-hatching is used to represent quantitative information, increasing magnitudes must be indicated by increasingly intense or dark types; if it is used only to distinguish one set of data from another according to some non-quantitative characteristic, any kinds of cross-hatching may be chosen that will afford the greatest possible contrast, usually by alternating light and dark types. It is possible to buy gummed paper printed in a great variety of cross-hatched patterns. This may be applied to the graph and trimmed to the desired shapes with great saving of time.

Contrast may also be achieved by differentiation in types of lines when several curves are being presented. A number of possible types are shown in Figure 47. The most important data, such as a combined index, can most effectively be represented by a solid black line, and the lines for the other curves should be selected so that they are easily distinguishable from one another and so that all are equally distinct. The curves representing the data should be heavier than the background lines on the graph. The usual order of these background lines is: border heaviest, followed by vertical and base scales and 100 per cent line, with other grid lines lightest.

FIGURE 47
TYPES OF LINES



When symbols are employed in a pictogram, boldness should be the primary consideration. One or two kinds of solid black symbols of standard shape, whose meaning cannot possibly be misunderstood are much more effective than a variety of outline sketches that cannot be interpreted without reference to some printed material.

The printing on a graph should be heavy enough to contribute to the effect of contrast, but not so heavy as to detract from the diagram itself. Vertical capital letters and figures with no ornamentation whatever are most suitable for this purpose, and are most easily read. The heaviest lettering will be used for the title, a smaller size for the legends, scales, etc., and probably the smallest of all for the reference to the source or other notes of explanation. It should go without saying that neatness in lettering is one of the most essential features of an artistic graph.

Technical Details

Certain techniques in graphic construction have come to be accepted as standard. These details should be observed, not because they are fixed by an arbitrary set of rules but because they are all founded upon the principle that graphs must give a clear idea with the minimum of effort on the part of the reader.

Title.—The title of a graph must meet the same requirements that were established for the title of a table.¹⁴ It should not give the conclusion to be drawn from the graph, as "Sales Larger This Year Than Last"; nor the method of analysis, as "Frequency Distribution of Number of Employees."¹⁵ Information such as the units of measure,

¹⁴ Chapter VIII, pp. 160–61.

¹⁵ Whenever a title of this sort is used in this text, it is because the method being illustrated is of greater importance to the reader than the actual data.

and the subgroups of a classification, which will be clearly indicated by the scales and legends or key, need not be included in the title of a graph.

Legend or Key.—These terms are often used interchangeably, but for this discussion “legend” will refer to labeling written within the bars, sectors, etc., or adjacent to the curves of time series to tell what each represents. (See the graphs in chapter XX.) “Key” will refer to a group of blocks or lines at the bottom of a graph indicating by sample pieces of the various lines or types of cross-hatching the significance of each wherever it may appear on the graph. (See the graphs in chapter XXII.) There is no hard-and-fast rule as to which method should be employed. In general, it may be said that if the lines or hatching are repeated in several different parts of the graph it is better to use one key that will apply to all parts. In a map where the same kind of hatching or dots appears in a number of different areas, a key is practically unavoidable. If the graph includes two or more circles or sets of bars, each having corresponding parts that follow a common system of shading, it is easier to follow one key than to read the same legends in several different places. On the other hand, if there is plenty of space to print a clear legend right next to the curve or within the sector, good judgment will indicate that this should be done.

Whenever legends are printed on the graph there are a number of points to consider. (*a*) There must be no possible confusion as to which curve or other part the legend is intended to mark. (*b*) If possible, avoid printing between any line or bar and the scale from which its value must be read. (*c*) On a closely crossed grid a white space inclosed in a border should be left clear for printing each legend. (*d*) Legends should be clearly printed, and worded as briefly as possible.

The use of a key likewise calls for some words of caution. (*a*) The key must be neatly ruled and adequately labeled. (*b*) The lines or hatching must correspond exactly to those used in the graph. (*c*) The key is a part of the graph and should be inclosed within the outer border, if the graph has a border; certainly it should never be transferred to some other page.

Scales.—A scale has two parts: its general label and the markings of its subdivisions.

Labels: The label states the unit of the vertical scale or the numer-

ical classification of the horizontal scale, as tons, dollars, years, etc. When the units are counted in large groups instead of singly, it also indicates the number in each group. To avoid confusion in locating the decimal point units should be grouped in thousands, millions, or billions, rather than in tens, hundreds, ten thousands, etc. For example, in a scale having a range of 0 to 2,000 tons, the values might well be written in full, or they might be shown as .25, .50, 1.00, 1.25, etc., under the label "tons in thousands" or "tons, 000 omitted." Just "000" means nothing; one does not know whether to interpret it as "in hundreds" or "000 omitted." Sometimes the ciphers omitted are stated in the title, in which case they should not be repeated in the scale label. The full contraction must be indicated in either one place or the other, never divided between the two.

If a multiple scale is being used, each scale must state the item to which it applies. A graph of index numbers or other per cents will be labeled either "index" or "per cent," with no reference to the original unit. The graphs shown in this text should be observed for standard practice in wording and arrangement. Note that the labels always read parallel to the base of the graph; that is, the label of the vertical scale appears across the top of that scale rather than vertically along the side, whereas for the horizontal scale it is in the center under the markings for years, months, etc.

Scale divisions and grid lines: Grid lines are scale divisions that are drawn all the way across a graph. They are usually fine solid lines of uniform thickness, although the lines indicating the ends of years, intervals of 50, etc., may be heavier than the intervening lines in order to set off the major divisions of a chart. Frequently only these main guide lines are drawn all the way across, the other values being indicated by short stubs along the axis. It is not necessary to indicate the numerical values of each one of these stubs but only enough of them to enable the reader to determine the value of any plotted point without too much trouble. The figures that are printed along a scale should be directly opposite the points to which they refer.

The methods of marking intervals in a time series require special attention in order to avoid confusion in reading the graph. First we shall consider the various ways of charting annual data. There are four alternatives, as shown in Figure 48.

In A, the year is indicated directly below the grid line on which the point is plotted. This is the preferable method for recording values

at the same date, such as May 1, for several successive years; it can also be used for yearly averages or totals, or whenever a single figure represents the entire year.

However, B is more suitable for the latter situation, since each point is plotted at the center of the space between two grid lines, the label for the year being also centered directly below it. Method B would also be correct for data as of June 30 or July 1, but not for any other given date, since the plotted points fall in the center of the yearly spaces.

Method C is incorrect for yearly averages or totals but could be used if the data were as of December 31 or January 1, provided this fact were made clear in the title. For any other data it is ambiguous because one cannot tell whether the year named in the center of the space applies to the point on the grid line preceding it or following it.

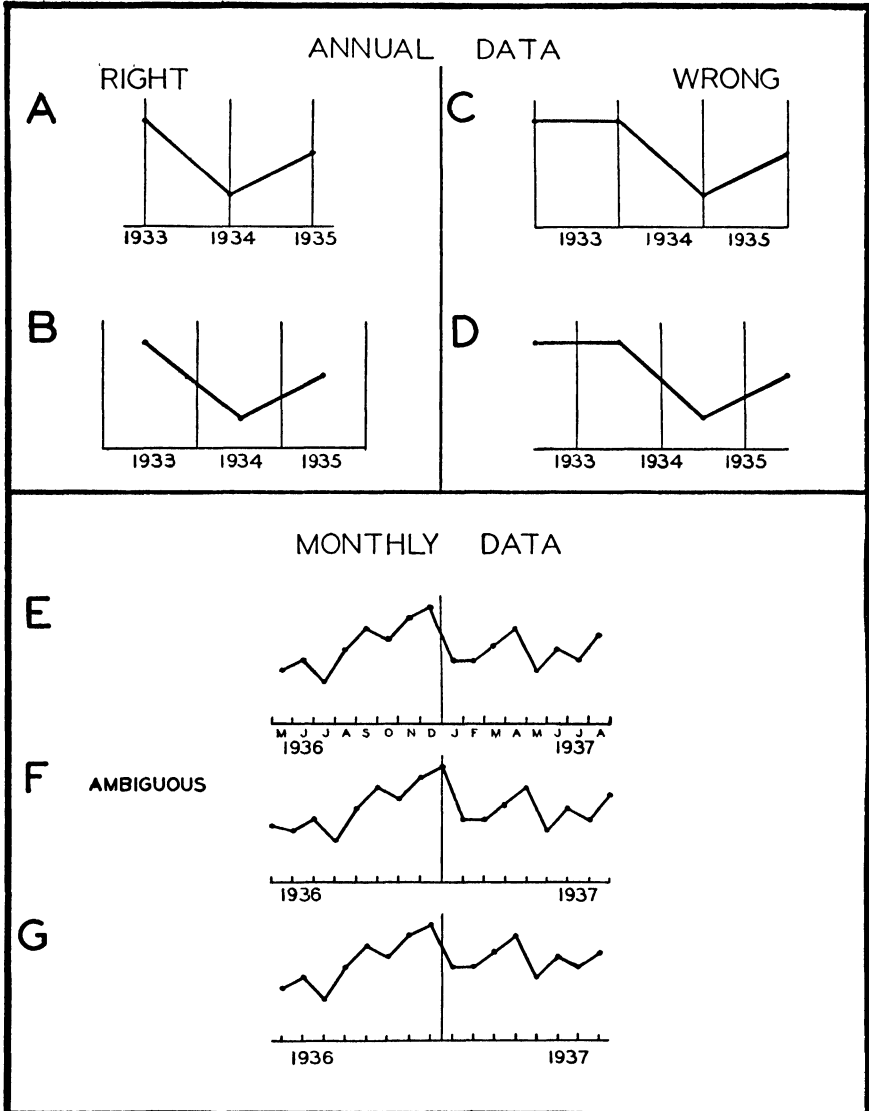
The last method, D, is the reverse of C; it is equally ambiguous and would be correct under no circumstances.

The same principles can be applied to the correct graphing of monthly data. The space representing a year is usually set off by grid lines. This annual space must therefore be divided into 12 equal parts, each of which represents one month. The spaces indicating the months need not be labeled in a long time series, although if the scale is large enough the abbreviation or initial of the month at the center of each space is an aid to the reader. Years should be printed out in full, horizontally below the monthly labels, at the center of the year's space.

If the monthly data are totals, averages, or mid-month recordings, they should be plotted at the center of each month's space. There will then be no value plotted directly on the grid line that marks the year's end, and it will be perfectly clear which point stands for December and which for January (See Figure 48-E). However, end-of-month data should be plotted at the end of each month's space, so that December 31 will quite correctly fall on the end-of-the-year's mark (See Figure 48-F). As in the case of annual data, Figure 48-C, this method is ambiguous unless the title of the graph indicates that the plotted points are recordings as of the last day of each month.

It would also be possible to locate monthly stubs at the center of each monthly interval, Figure 48-G. This corresponds to method A for annual data, and causes no difficulty in reading the graph. However, since the years are labeled at the center of the year's space, it

FIGURE 48
METHODS OF PLOTTING TIME PERIODS



is more consistent to plot each month at the center of a space as in 48-E, rather than at a stub.

Accompanying Table.—Since no graph aims to record exact numerical values it is always desirable to provide an accompanying table for the benefit of the reader who wishes to verify or make further use of

the information. The table should appear on the same page as the graph, or on the page facing it, and both should read in the same direction. Students seldom realize the unfavorable impression made on an instructor or any other reader who is forced to compare a graph that reads vertically with a table that reads horizontally, or vice versa. It has already been suggested in the discussion of balance that a brief table may be printed on the graph in some unoccupied space if it does not interfere with the graphic presentation.

Reference and Notes.—The necessity for quoting the source and noting any discrepancies in the data was explained in discussing the requirements for statistical tables.¹⁶ Practically the same rules may be applied to graphs although, if the information has been given in an adjacent table, reference on the graph need only be made to that table as the source. The reference, either to the accompanying table or the original source, is usually printed in the lower right-hand corner of the graph.

Important Points in Actual Construction.—No attempt will be made in this text to give a summary of the principles of mechanical drawing. A course in that subject is a great aid to anyone who wishes to draw graphs neatly and correctly. It is possible, however, to secure manuals on the subject, lettering guides, etc. A study of the instructions that come with ruling and lettering pens should help the student in his first efforts to use India ink. With a few hours of practice anyone can learn to handle a ruling pen and lettering stencils without blotting. Accuracy in scale and angle measurement is not beyond the capacity of the average person. Even lettering by hand is only a matter of a little care and practice in copying from lettering guides.

In outline form, the order of steps in drawing a graph is as follows:

- a) Check all data for accuracy in computation or in copying from source.
- b) Plan the scales to conform to the correct size and proportion, within the range of the data.
- c) Measure scales and draw axes and guide lines in pencil. (More pencil guide lines will be needed than finally appear on the graph.)
- d) Plot the data.

¹⁶ Chapter VIII, p. 162.

- e) Check plotting, reading from the points back to the data.
- f) Plan spacing of lettering—titles, scales, labels, key, source, etc.
- g) Ink in all lines, including borders, guide lines, etc., taking time to let each section dry before doing further work near it.
- h) Ink in lettering, using stencils if possible.
- i) Erase all pencil marks.

PROBLEMS

1. What is the order of arrangement of the bars in the bar charts appearing in chapter XIII?
2. The following is the production of anthracite coal in the United States at five-year intervals from 1900–40 (thousand short tons):

YEAR	PRODUCTION	YEAR	PRODUCTION
1900	57,468	1925	61,817
1905	77,660	1930	69,385
1910	84,485	1935	52,159
1915	88,995	1940	50,052
1920	89,598		

- a) Present these data in a bar diagram.
- b) Why is this form superior to a line diagram for these data?
- c) How would you read from this diagram the 40-year history of the anthracite coal industry?
3. Find an applied use of the band chart in a published source. Describe the contents of the chart and state briefly the major relations portrayed.
4. a) Plot the following data on four separate charts, corresponding to the four methods shown in Figure 44, A, B, C and D, pages 328–29. Use 1935 as the base for the index number graph.
b) Explain briefly what you think each graph shows.

APPROXIMATE SALES, GROSS PROFIT AND NET PROFIT OF A SMALL MANUFACTURING CONCERN, 1932 TO 1938

YEAR	SALES	GROSS PROFIT	NET PROFIT
1932	\$15,000	\$1,000	\$100
1933	22,000	3,000	400
1934	18,000	1,500	200
1935	26,000	4,500	800
1936	20,000	2,250	400
1937	42,000	6,750	1,600
1938	33,000	3,375	800

5. a) Draw a graph of the grapefruit production data given below.
- b) Study and interpret the facts shown by your graph.

PRODUCTION OF GRAPEFRUIT IN THE UNITED STATES, 1919 TO 1939 *

YEAR	PRODUCTION (Million Boxes)			
	California	Florida	Texas	Total
1919	.4	5.9	...	6.3
1924	.4	8.8	.2	9.4
1929	1	8	2	11
1934	2	15	3	20
1935	2	12	3	17
1936	2	18	10	30
1937	2	15	12	29
1938	2	24	16	42
1939	2	17	15	34

* *Agricultural Statistics*, 1938; and *Crops and Markets*, December, 1939.

6. a) Why should every ordinary scale chart have a zero base line?
- b) Why in using two vertical scales on the same chart should the two scales bear some fixed relation to each other?
- c) Under what conditions is it justifiable to use colored inks in drawing charts?
- d) In a two-dimensional graph how do you determine which values to plot on the base scale?
7. a) Find one published graph that you consider is correctly and effectively drawn, and explain why you think so.
- b) Find one published graph that you think has certain features that are incorrect, and give reasons.
- c) Find one published graph that you consider ineffective, and suggest changes that might add to its effectiveness.

REFERENCES

- ARKIN, HERBERT, and COLTON, RAYMOND R., *Graphs, How To Make and Use Them*. New York: Harper & Bros., 1936.
- BRINTON, WILLARD C., *Graphic Presentation*. New York: Brinton Associates, 1939.
- HASKELL, ALLAN C., *Graphic Charts in Business*. New York: Codex Book Company, 1922.
- KNOEPPPEL, CHARLES E., *Graphic Production Control*. New York: The Engineering Magazine Co., 1920.
- LEHOCZKY, PAUL N., *Alignment Charts, Their Construction and Use*, Engineering Experiment Station Circular No. 34. Columbus, Ohio: The Ohio State University Studies, 1936.

- MODLEY, RUDOLF, *How To Use Pictorial Statistics*. New York: Harper & Bros., 1937.
- MUDGETT, BRUCE D., *Statistical Tables and Graphs*. Boston: Houghton Mifflin Co., 1930.
- RIGGLEMAN, JOHN R., and FRISBEE, IRA N., *Business Statistics*. New York: McGraw-Hill Book Co., Inc., 1932, Appendix III.
- RIGGLEMAN, JOHN R., *Graphic Methods for Presenting Business Statistics*. New York: McGraw-Hill Book Co., Inc., 1926.
- SMITH, HERBERT G., *Figuring with Graphs and Scales*. Stanford University, California: Stanford University Press, 1938.
- Time Series Charts, A Manual of Design and Construction*. New York: The American Society of Mechanical Engineers, 1938.

CHAPTER XV

FREQUENCY DISTRIBUTIONS AND GRAPHS

FREQUENCY DISTRIBUTIONS

A FREQUENCY distribution is simply one of the methods of classification of data, and in form resembles any other statistical table. An example that has already been introduced in the text is Table 16-A, chapter VIII (wage rates of explosives workers). This particular form of classification has been reserved for special treatment because the idea of grouping large masses of data according to their quantitative characteristics is one of the most fundamental processes in statistics. In many phases of business operations it is an important first step toward more advanced analysis.

A frequency distribution is always a classification of data in which the items are combined in groups according to size. The "ordered classification" is the independent variable, and the numbers of items that appear in the several groups become the dependent variable. For example, the dependent variable, number of firms, might be classified in groups according to annual dollars of sales, number of tons of product shipped weekly, number of employees, or hourly wage rates paid. On the other hand various dependent variables might be tabulated with any of these classifications (independent variables); e.g., with a wage classification the dependent variable might be either the numbers of employees receiving the various rates, the numbers of years in which the several wage rates were paid, or the numbers of states in which the rates were standard. The number of units, or items counted, in each group is called its *frequency*.

According to this method of grouping large numbers of detailed observations, each individual item or occurrence loses its identity and becomes one of a larger group that has a broader definition of quantity or value. For instance, in grouping wage data a single wage payment of \$18.52 might become one of a group of 65 payments designated as "\$18.00 to \$18.99." It follows that the basic requirements for a satisfactory frequency distribution are: (1) the value of each individual item must be known at the outset, and (2) the values must be grouped

in such a way that the summary table will accurately represent the individual items from which it is compiled.

First Steps in Analysis

The steps that are followed in making an analysis by means of a frequency distribution will be illustrated by rent data that were collected in Columbus, Ohio. This small but representative sample of 155 rent payments was secured as a by-product of a study of consumer habits in the patronage of dry-cleaning establishments.

Arraying the Data.—The initial step was to list each rent payment as the reports came in from the interviewers. The result is shown in Table 56. This random listing gives no clue whatever to any possible

TABLE 56
RENTS PAID BY 155 FAMILIES IN A CONSUMER
SURVEY IN COLUMBUS, OHIO
(Dollars Per Month)

\$50	\$ 8	\$60	\$50	\$75	\$21
25	80	75	50	75	25
18	9	75	50	53	25
75	15	95	60	55	15
55	16	35	60	60	18
53	35	35	25	60	80
30	35	35	24	60	75
50	50	35	35	35	51
31	30	32	25	65	75
24	28	22	25	60	51
13	27	22	25	40	50
15	25	20	30	40	45
40	40	40	40	45	35
65	20	9	12	35	35
68	18	17	13	35	35
70	16	18	15	25	30
80	13	30	18	15	30
80	85	30	30	16	30
35	90	35	35	18	18
35	80	35	30	21	20
40	65	35	40	21	30
45	35	85	40	30	35
40	51	95	40	30	35
40	51	80	40	35	35
48	60	70	85	40	35
50	50	35	35	35	

interpretation of the data. The question now arises, if they were ranged in order of size would any significant relationship appear? In order to answer this they were next arranged in an array, as shown in Table 57.

TABLE 57
 ARRAY OF RENTS PAID BY 155 FAMILIES
 IN A CONSUMER SURVEY IN COLUMBUS, OHIO
 (Dollars Per Month)

\$ 8	\$21	\$30	\$35	\$48	\$65
9	21	30	35	50	65
9	21	30	35	50	65
12	22	30	35	50	68
13	22	30	35	50	70
13	24	31	35	50	70
13	24	32	35	50	75
15	25	35	35	50	75
15	25	35	35	50	75
15	25	35	40	50	75
15	25	35	40	51	75
15	25	35	40	51	75
16	25	35	40	51	75
16	25	35	40	51	80
16	25	35	40	53	80
17	25	35	40	53	80
18	27	35	40	55	80
18	28	35	40	55	80
18	30	35	40	60	80
18	30	35	40	60	85
18	30	35	40	60	85
18	30	35	40	60	85
18	30	35	40	60	90
20	30	35	45	60	95
20	30	35	45	60	95
20	30	35	45	60	

There are various ways of putting data in an array, depending somewhat on the form in which they have been collected. If each item is on a separate card or sheet or schedule, these could first be sorted according to size and then listed. Or they might be tallied by assigning one line of a ruled sheet of paper to each possible value and then writing down each rent as it appears from the random assortment. The result would appear as in Figure 49. If the rents or tally marks are evenly spaced the resulting rows take the place of a rough bar diagram in indicating the distribution of frequencies according to rental value. To get a true picture it is necessary to have a line for every unit rental value in the series, whether or not it has any frequencies.

An alternative method for analyzing either the sorted or tallied data would be to draw a simple bar diagram as shown in Figure 50. The range of values is clearly revealed by comparing the shortest bar at the top of the graph with the longest one at the bottom. The values at which there are concentrations and the number of similar items of various values can be seen by looking for the bars of equal length. This type of graph is seldom used for final presentation unless it portrays characteristics which are peculiar to the data and which can-

FIGURE 49
TALLY OF MONTHLY RENTS PAID BY 155 FAMILIES IN A CONSUMER SURVEY IN

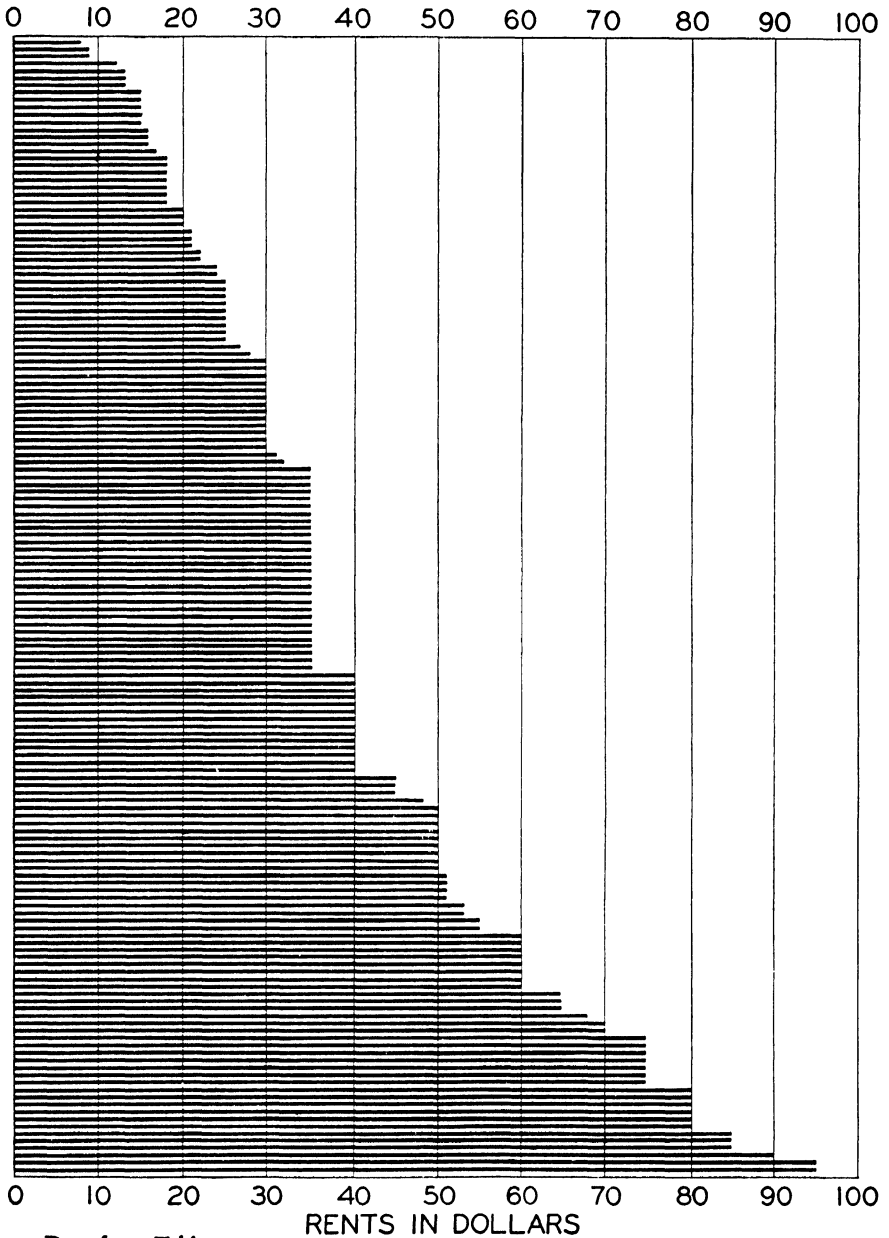
COLUMBUS, OHIO			
RENT	NUMBER OF FAMILIES	RENT	NUMBER OF FAMILIES
\$ 8	I	\$52	
9	II	53	II
10		54	
11		55	II
12	I	56	
13	III	57	
14		58	
15	III	59	
16	III	60	III III
17	I	61	
18	III II	62	
19		63	
20	III	64	
21	III	65	III
22	II	66	
23		67	
24	II	68	I
25	III IIII	69	
26		70	II
27	I	71	
28	I	72	
29		73	
30	III III III	74	
31	I	75	III II
32	I	76	
33		77	
34		78	
35	III III III III III III	79	
36		80	III I
37		81	
38		82	
39		83	
40	III III IIII	84	
41		85	III
42		86	
43		87	
44		88	
45	III	89	
46		90	I
47		91	
48	I	92	
49		93	
50	III IIII	94	
51	IIII	95	II

not easily be shown by any other graphic form. It is a helpful graph in preliminary analysis, however, for it provides the basis for the examination which is necessary before the data can be grouped.

Preliminary Grouping of the Data.—From a study either of Figure 49 or Figure 50 it can be seen at once that the whole range of data extends from a low value of \$8 to a high of \$95. The largest number of rents appears to be at \$35 but there are certain other concentration points, notably at \$25, \$30, \$40, \$50, and \$60.

FIGURE 50

ARRAY OF RENTS PAID BY 155 FAMILIES IN COLUMBUS, OHIO
(Each bar represents one family)



Data from Table 57.

At individual values: This suggests the next logical step which is the initial grouping process, that is, to count all the items having the same value. In the frequency array shown in Table 58 all of the rents appear in order along with the number of times each individual rent occurs.¹

TABLE 58
FREQUENCY ARRAY OF MONTHLY RENTS PAID BY 155 FAMILIES IN A
CONSUMER SURVEY IN COLUMBUS, OHIO

RENTALS PAID	NUMBER	RENTALS PAID	NUMBER	RENTALS PAID	NUMBER
\$ 8	1	\$25	9	\$53	2
9	2	27	1	55	2
12	1	28	1	60	8
13	3	30	13	65	3
15	5	31	1	68	1
16	3	32	1	70	2
17	1	35	28	75	7
18	7	40	14	80	6
20	3	45	3	85	3
21	3	48	1	90	1
22	2	50	9	95	2
24	2	51	4	Total	155

The characteristics of the data begin to stand out more clearly. We now know exactly how many rents of each amount were paid. The \$35 rent occurs 28 times, having the highest frequency in the array, while the \$30 and \$40 amounts are almost tied for second place with 13 and 14 frequencies, respectively. The rents less than \$35 are concentrated between \$8 and \$32, whereas those greater than \$35 are spread over a range from \$40 to \$95.

In class intervals: However, there are still too many separate values listed for easy comprehension of the complete information regarding these rents. The entire situation can be readily grasped only after the 155 items have been grouped into a few classes. These classes must cover the entire range from \$8 to \$95 and must represent as far as possible the characteristics that have been observed from studying the individual items. Before continuing this process with the rent data it will be necessary to consider a number of points that must always be taken into account in determining the groups of a frequency distribution./

Principles for Grouping Data

The questions that must be answered before deciding how to group any individual data are:

1. Into how many groups, or class intervals, should a given set of data be divided?
2. What should be the width of each interval?
3. At what values should the class limits be set?
4. How should the class limits be designated?

Number of Intervals.—The number of intervals is less important than the width of intervals and the values of class limits. The exact number used will finally be determined by the range of the data after these other two points have been decided. There are a few rule-of-thumb guides, however, which aid in roughly determining the number of intervals. In the first place, since the purpose of grouping is to aid in the summary and comprehension of data, there should be no more intervals than can be quickly grasped. In the second place, the number of intervals cannot be so small that important characteristics of the data are concealed. These two criteria, however, are rather general to serve as operating rules in determining how many class intervals to use in a given frequency distribution.

Statisticians have indicated the number of intervals which in general meet the requirements of distributions of most kinds of data. Yule says that desirable conditions will usually be fulfilled if the "number of classes lies between 15 and 25."¹ A minimum is suggested by the statement that it is "desirable to have more than eight classes."² It has been suggested that the number of classes can be determined by the use of a formula which has been developed from the theory of binomial expansion. The formula as developed by Sturges³ is: Number of class intervals = $1 + 3.322 \log$ of number of observations in the distribution. Solution of this formula indicates the following number of class intervals should be used with designated numbers of observations:

NUMBER OF OBSERVATIONS	NUMBER OF CLASS INTERVALS
100	8
200	9
400	10
600	10
800	11

The number of classes should be determined only after making a

¹ G. Udny Yule and M. G. Kendall, *An Introduction to the Theory of Statistics* (London: Charles Griffin and Co., Ltd., 1937), p. 85.

² Frederick E. Croxton and Dudley J. Cowden, *Practical Business Statistics* (New York: Prentice-Hall, Inc., 1937), p. 153.

³ H. A. Sturges, "The Choice of a Class Interval," *Journal of the American Statistical Association*, Vol. XXI (1926), pp. 65-66. Cf. Harold T. Davis and W. F. C. Nelson *Elements of Statistics* (Bloomington, Indiana: The Principia Press, 1935), p. 16.

careful study of all the characteristics of the data, instead of applying this formula indiscriminately.

Width of intervals.—There are no arbitrary criteria for determining the width of the class intervals in any distribution, but the following considerations are pertinent to the problem.

1. Class intervals should not be so wide that too much of the detail of the distribution is lost through grouping. It is true that the purpose of the frequency distribution is to summarize and to reduce the volume of the data to workable proportions, but the features of the data should not be concealed or eliminated through grouping in wide intervals.

2. Little is gained on the other hand by arranging data in very small intervals, if the number of classes then remains too large to provide an effective summary.¹

3. The total number of frequencies in the distribution serves as a rough guide to the size of the class intervals⁴ that should be employed. That is, when there are a great many observations the intervals can be relatively small because it will be permissible to have a large number of classes. Within the same range of data, if there are only a few observations the number of class intervals must be smaller and the width of the intervals will be correspondingly greater.

4. If there is any discernible pattern in the distribution, however, it will serve as a much better guide to the size of the classes. For instance, if hourly wage rates are being studied, it may be found that more men are paid even five- or ten-cent rates than any intervening amounts. This pattern must be preserved in the grouped data through correct choice of the size of class intervals. The class width should be five cents or some multiple of five cents so that there will be an equal number of concentration points within each interval.

5. The class intervals should be chosen so that there will be a minimum number of classes that contain no frequencies.

6. If the distribution which is being constructed is to be compared with others that are already prepared, the intervals in the new distribution should be made to conform with those in the previous distribution. If several different but comparable distributions are being prepared at the same time, the size of the class intervals must be

⁴ Related to the formula on page 356, Sturges recommends the following formula for the determination of the size of class intervals:

$$\text{Size of class intervals} = \frac{\text{range of data}}{1 + 3.322 \log \text{ of number of observations}}$$

established in view of the characteristics of the several distributions.

7. In a given distribution every effort should be exerted to make all class intervals equal. The distribution of data in unequal intervals makes analysis difficult and certain kinds of computation impossible. Although unequal class intervals should not ordinarily be used, there are certain cases in which they are unavoidable.

a.) In some cases where a few high-valued observations are widely dispersed, they may be grouped in increasingly large class intervals, so as not to reveal the identity of the individual cases. Unequal class intervals are frequently employed for this reason by various government departments. Table 59-A indicates a frequency distribution of this type.

b.) Analysis of the data may indicate that unequal class intervals define the homogeneity of the observations more accurately than equal intervals. In Table 59-B for instance, unequal class intervals were used because the management felt that these divisions of purchases gave them the most assistance in their merchandising plans.

c.) Equal relative increases in the widths of class intervals may be of more significance to a particular distribution than equal absolute changes. Consequently, a frequency distribution which appears to have unequal class intervals may in reality have intervals which are increasing in size at a uniform rate.

TABLE 59

A		B	
NUMBER OF BROADCASTING STATIONS IN THE UNITED STATES, BY ANNUAL REVENUE RECEIVED, 1935 *		NUMBER OF PURCHASERS AT THE COLUMBUS CONSUMERS' COOPERATIVE ASSOCIATION, BY VALUE OF PURCHASES, JULY 1, 1937, TO DECEMBER 31, 1937 †	
ANNUAL REVENUE	NUMBER OF STATIONS	VALUE OF PURCHASES	NUMBER OF PURCHASERS
Less than \$10,000	48	\$ 0.00 to \$19.99	248
\$10,000- 24,999	67	20.00 to 39.99	140
25,000- 49,999	59	40.00 to 89.99	202
50,000- 99,999	46	90.00 to 149.99	74
100,000-249,999	45	150.00 to 299.99	49
250,000-499,999	17	300.00 and over	11
500,000 and over	7		
Total	289	Total	724

* Radio Broadcasting, *Census of Business*: 1935, p. 53.

† Unpublished Study of Patron Purchasing of the Columbus Consumers' Cooperative Association, 1938.

8. Finally, the problem of determining the size of the class intervals in a frequency distribution cannot be separated from that of establish-

ing the location of the class limits, that is, the values at which each group in the distribution will begin and end. It is usually necessary to consider these two points together before arriving at any decision with regard to either of them.

Limits of Intervals.—As in the determination of the size of class intervals, there is no single guide to the location of class interval limits. The several general criteria which follow indicate the nature of the problems that arise and the solutions that may be employed in various types of distributions.

1. If there is no pattern nor other guide, the very simple procedure of dividing the range of the data in the distribution by the approximate number of intervals may be used. For convenience, the actual dividing points thus established would then be rounded to the nearest whole numbers. Such a procedure, however, may completely disregard important characteristics of the data that should be revealed by the distribution, and should never be employed unless a careful study of the data has failed to reveal any pattern.

2. If a pattern is discovered, the limits should be set so as to preserve in each group the characteristics of the individual items the same as in determining the width of the intervals. This can be done by observing the values at which the frequencies are greatest and establishing the class limits so that these values fall at the midpoints. For example, in the case quoted of concentration of wage rates at five-cent intervals, the limits would not be set at 25, 30, 35 cents, etc., but at 27.5, 32.5, 37.5, etc., so that the concentration point falls at the center of each interval. If ten-cent intervals were used the limits would be set at 27.5, 37.5, etc., or 22.5, 32.5, etc., so that there would be two concentration points in each interval, each equidistant from the center and from either end.

3. Even though no special pattern is present, the class limits should be established so that the value half way between the class limits approximates the arithmetic average of the observations included in each class interval. This midvalue of each interval is called the "midpoint" or the "class mark."

4. When possible, the limits should be chosen so that the midpoints are integers. The importance of this guide will become clear in the computation of averages from frequency distributions. As in the case just cited, it is usually more important to have the midpoint an integer than to have the class limits themselves integers.

5. 'On some occasions one or both ends of the distribution may be left "open"; the minimum and maximum values are not shown,' (See Table 59.) These open-end frequency distributions are sometimes necessary to conceal the identity of the cases at the extremes, but the absence of limiting values is a serious handicap in subsequent analysis.

Designation of Class Limits.—The interpretation of the data in a frequency distribution and the evaluation of their accuracy depend largely upon the precise designation of the class limits. The method of designation may in turn depend upon the nature of the data involved.

Discrete or continuous data: An important consideration is whether the data are discrete or continuous. *Discrete data* are those which occur only at exact values at regular intervals but never at any intervening values. For example, stock prices are quoted in eighths of a point. With the exception of a few special listings, no stock would be quoted, at any price between $\frac{3}{8}$ and $\frac{1}{2}$, $\frac{1}{8}$ and $\frac{3}{4}$, etc. Likewise a classification according to number of employees could never be anything except whole numbers. In the latter case the classes would naturally be designated as 1 to 10, 11 to 20, etc., and no question could arise as to any fractional value between 10 and 11.

/ Continuous data, on the other hand, are those which may occur at every conceivable point along a continuous scale of values. This distinction between measured values and separate items and the methods for handling each statistically will be discussed in greater detail in chapter XVII. As a matter of fact, a classification in discrete units is much more puzzling to handle correctly in computation but, when class limits are being designated, continuous data afford a greater variety of alternative methods.

Examples of methods: Some of the methods for designating class limits are better than others in clarifying the actual limits that are employed. One of the most common methods has been shown in Table 59. Other methods are illustrated in Figure 51.

Of the four methods, the one shown in Figure 51-A is the poorest for it may be ambiguous. It is not clear whether exactly \$500 is included in the first or the second interval. In spite of this weakness, this form is widely used and is ordinarily interpreted as \$250 and under \$500.

Figure 51-C differs from the form in Table 59-A only in the loca-

FIGURE 51
METHODS OF DESIGNATING CLASS LIMITS

A	B	C
\$ 250-\$ 500	Under 30 cents	\$ 51-\$100
500- 750	30 and under 35 cents	101- 150
750- 1,000	35 and under 40 cents	151- 200
1,000- 1,250	40 and under 45 cents	201- 250
1,250- 1,500	45 and under 50 cents	251- 300
1,500- 1,750	50 and under 55 cents	301- 350
1,750- 2,000	55 and under 60 cents	351- 400

tion of the class limits: in one the round hundreds and fifties are the *upper* values of the classes, whereas in the other various multiples of round thousands are the *lower* values of the classes. Either of these forms can be used, Figure 51-C being preferred for discrete data, and Table 59-A for continuous data. The purpose in each case is to make the class intervals equal. Discrete data, which usually start at a value of one, must read 1-50, 51-100, etc., whereas continuous data are measured from zero and read 0-99, 100-199, etc.

In the case of continuous data, the person who prepares the table must make a decision regarding significant figures, whether he uses the form in Table 59-A, 59-B, or even Figure 51-B. In the first case the values are rounded at dollars, so that presumably any value up to \$9,999.50 would go in the first class, and \$9,999.50 and over in the second, etc. Similarly in Table 59-B the dividing line is \$19.995.

The designation in Figure 51-B indicates an indefinite number of decimal places, although in actual practice the dividing point between classes would seldom be carried any farther than a half cent.

Another common method of indicating class values, especially when the intervals are quite small, is by the value of the midpoint, as average grade, 75, 80, 85, 90, where 80 per cent includes everything from 77.5 to less than 82.5, etc. In other cases, classes that are listed in this way may represent single unit values of a discrete series.

Of all these possible methods for designating class limits, Figure 51-B is the least ambiguous. The units in which any particular data are expressed will ordinarily make clear to the reader to how many significant figures the class limits have been carried. This method requires more space for the stub of the table but it is nevertheless the method preferred by the authors.

An Example of the Preparation of a Frequency Table

The principles set forth in the preceding section will now be applied in the preparation of a frequency distribution of the sample of rents

paid in Columbus, Ohio. The preliminary preparation of these data was carried out earlier in the chapter, leaving them in the form found in Table 58. The next step is to determine the number of class intervals, the width of interval and the interval limits that will produce a concise and effective table. This means a table that is compact enough to be grasped quickly, and with the details arranged so that no essential characteristic of the data will be lost.

The range of \$87 between the lowest and highest rents paid immediately suggests the use of nine \$10 intervals. Nine intervals for 155 items appears reasonable and the \$10 width is convenient. The most important consideration, however, is the existence in the distribution of concentration points at the \$5 and \$10 rents, i.e., the tendency to fix rents at \$25, \$30, \$35, \$40, etc. This means that the width of the class interval must be \$5 or some multiple of \$5. A \$5 interval would give too many classes, a \$15 interval too few, hence \$10 emerges as the proper width to use and the number of intervals is automatically set at nine, if the first class is set at \$7.50 to \$17.50. The first class might also read \$2.50-\$12.50, so that the last of ten classes would read \$92.50-\$102.50. There is no general rule that requires the use of either one or the other of these systems of intervals and the distribution of frequencies will, of course, be different according to which is employed. Perhaps the best plan is to regroup the initial \$5 intervals in \$10 intervals according to both systems and then select the one that gives the smoother distribution or appears to be the better description of the data. In the distribution of rents the \$7.50-\$17.50 set of intervals seems preferable.

The same circumstance that led to the selection of \$10 intervals also becomes the guide to the proper class limits. Each class will contain two points of rent concentration. These must fall at equal distances from the center and ends of the class in order to meet the requirement that the average value of the items included in any class shall be approximately equal to the midpoint of that class. Hence the first class containing the \$10 and \$15 concentration points must have its midpoint at \$12.50. The next containing the \$20 and \$25 concentration points must have its midpoint at \$22.50 and so on. The class limits, therefore, must be \$7.50, \$17.50, \$27.50, . . . \$97.50.

Actual Frequencies.—The three parts of Table 60 contain different distributions resulting from the use of three distinct sets of \$10 class intervals. That is, A, B, and C are independent distributions each

TABLE 60

THREE FREQUENCY DISTRIBUTIONS OF MONTHLY RENTS PAID BY 155 FAMILIES
IN A CONSUMER SURVEY IN COLUMBUS, OHIO

CLASS INTERVAL	(1) FREQUENCY	(2) CLASS MARK	(3) CLASS INTERVAL AVERAGE
Frequency Distribution A			
\$ 5 and under \$ 15.....	7	\$10	\$11.0
15 and under 25.....	26	20	18.5
25 and under 35.....	26	30	28.2
35 and under 45.....	42	40	36.7
45 and under 55.....	19	50	49.6
55 and under 65.....	10	60	59.0
65 and under 75.....	6	70	67.2
75 and under 85.....	13	80	77.3
85 and under 95.....	4	90	86.2
95 and under 105.....	2	100	95.0
Total	155		
Frequency Distribution B			
\$ 0 and under \$ 10.....	3	\$ 5	\$ 8.7
10 and under 20.....	20	15	15.8
20 and under 30.....	21	25	23.6
30 and under 40.....	43	35	33.3
40 and under 50.....	18	45	41.3
50 and under 60.....	17	55	51.2
60 and under 70.....	12	65	61.9
70 and under 80.....	9	75	73.9
80 and under 90.....	9	85	81.7
90 and under 100.....	3	95	93.3
Total	155		
Frequency Distribution C			
\$ 7.50 and under \$17.50.....	16	\$12.50	\$13.6
17.50 and under 27.50.....	27	22.50	22.0
27.50 and under 37.50.....	44	32.50	33.2
37.50 and under 47.50.....	17	42.50	40.9
47.50 and under 57.50.....	18	52.50	51.0
57.50 and under 67.50.....	11	62.50	61.4
67.50 and under 77.50.....	10	72.50	73.3
77.50 and under 87.50.....	9	82.50	81.7
87.50 and under 97.50.....	3	92.50	93.3
Total	155		

of which has been constructed from Table 58. Column 1 records the number of rents falling within the limits indicated for the several classes in each of the three distributions. In distributions A and B the two concentration points fall at the beginning and center of the intervals. In distribution C, however, the center of the interval or class mark lies midway between the two concentration points.

Columns 2 and 3 have been added to Table 60 to demonstrate the superiority of distribution C. The class marks in column 2 con-

form to the definition previously given. The class averages in column 3 are obtained from the array in Table 57. For example the seven rents recorded between \$5 and \$15 total \$77 or an average of \$11. Parallel computations for each class of each distribution lead to the averages as recorded. Comparison of column 3 with column 2 shows that in distribution A the averages are less than the class marks in all classes except the first and that the differences are appreciable except in the "\$45 and under \$55" class. Likewise, in distribution B the averages are below the class marks except in the first and second classes. In distribution C four averages are above and five below their respective class marks and the differences between the averages and the class marks are small. Distribution C, therefore, meets the requirement that class marks should approximate the actual averages of the items included, whereas the other two distributions contain a definite bias. This bias would have an adverse effect upon any numerical measures computed from those distributions.

If the size of the rent sample were increased to several thousand items, the averages and class marks in distribution C would tend to coincide, but the bias would persist in distributions A and B. For this reason the characteristics of the universe "rents in Columbus, Ohio" can be studied from distribution C only.

Percentage Frequencies.—Percentage frequencies are preferable to actual frequencies for some purposes. Table 61 (which has been set up with title and headings such as would be used in a presentation table, in contrast to the work-table headings of Table 60) shows both the actual frequencies from Table 60-C, and their percentage distribution. Two major uses of percentage frequencies should be mentioned: (1) the comparison of the individual frequencies with each other and with the total, and (2) comparisons between two or more distributions having the same or equivalent class intervals. Thus from Table 61, column 2, it is apparent that more than one-fourth of the rents were between \$27.50 and \$37.50, that less than one-fourth were above \$57.50 and that more than one-fourth were less than \$27.50. The advantages of the percentage frequencies in comparing two distributions graphically will be shown later in the chapter.

GRAPHS OF FREQUENCY DISTRIBUTIONS

The frequency distribution is a separation of a whole into parts, the frequencies being merely a record of the number of individual

TABLE 61
 NUMBER AND PERCENTAGE DISTRIBUTION OF FAMILIES IN COLUMBUS, OHIO,
 ACCORDING TO VALUE OF MONTHLY RENTALS PAID

RENTALS PAID	FAMILIES	
	(1) Number	(2) Percentage Distribution
\$ 7.50 and under \$17.50.....	16	10.3
17.50 and under 27.50.....	27	17.4
27.50 and under 37.50.....	44	28.4
37.50 and under 47.50.....	17	11.0
47.50 and under 57.50.....	18	11.6
57.50 and under 67.50.....	11	7.1
67.50 and under 77.50.....	10	6.4
77.50 and under 87.50.....	9	5.8
87.50 and under 97.50.....	3	1.9
Total	155	100.

items falling in each quantitative class of the distribution. The major purpose in presenting a distribution graphically is to emphasize the relation of the parts to the total and to each other.

The graphic methods used in presenting frequency distributions are perhaps more standardized than in the case of any other kind of statistical data. The forms have become so widely accepted that it is necessary to follow without noticeable deviation the generally accepted rules for their construction. This does not imply, however, that we may not inquire into the underlying principles that have led to the development and universal acceptance of these methods.

Construction and General Characteristics

As was indicated at the end of chapter XIII, frequency distribution graphs are of the two-dimensional variety. The class intervals are always plotted on the horizontal axis and the frequencies on the vertical axis. Ordinary arithmetic scales are used on both, except for some very specialized types of distribution which are excluded from the present discussion. The vertical scale must always begin at zero, but the horizontal scale need include only the range of the class values, plus an extra interval at either end. The two most common frequency diagrams are the histogram and the frequency polygon.⁵

⁵ A third diagram, the smooth curve, is often discussed along with the histogram and the frequency polygon. It is a trace of the form which the frequency distribution would take if a very large number of cases were included and class intervals become infinitesimally small. From the point of view of universe and sample this concept is of considerable importance and will be employed in chapter XXVIII. The smooth curve has little application at the elementary level; hence no further reference will be made to it in this chapter

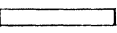
The Histogram.—This form of diagram consists of contiguous rectangles, or columns, ranged along the base scale, the height of each one being determined by the number of frequencies in the class upon which it stands. The total combined area of all the columns represents the total number of frequencies in the distribution. It may be considered that each column is like a pile of coins, each coin representing a single frequency. The thickness of the coin equals the value of one frequency on the vertical scale, and its diameter corresponds to the width of the class interval. Viewed from the front each coin occupies a narrow rectangular space . Several such adjacent piles of different heights, would look very much like a frequency histogram. If a few coins were moved from one pile to another the total front view area, representing the total number of coins or frequencies, would remain the same regardless of changes in the distribution.

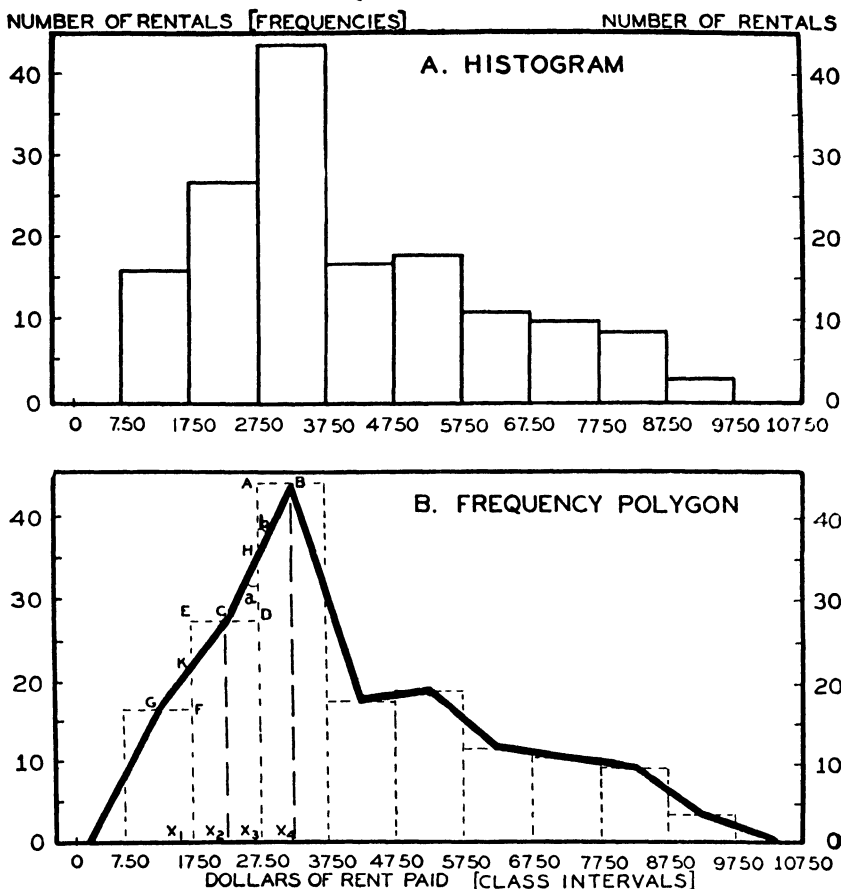
Figure 52-A which represents the rent data distribution of Table 60-C, illustrates the important features of all histograms. The greatest concentration of frequencies is at once apparent from the location of the tallest column on the base \$27.50 to \$37.50. The other columns start from zero frequency on the left and gradually increase in height as they approach this class of maximum frequency, while those on the right fall away from it and finally reach zero again. This is the characteristic shape of a frequency graph portraying the chance occurrence of a set of homogeneous events. Variations from this usual shape will be discussed later in the chapter.

/ Another point to be noted from the histogram is that each column rests not upon a single point but upon the entire interval included within the class limits. This indicates that the frequencies in any interval are spread over that interval, and that the base scale values occur in a continuous sequence. /

The Frequency Polygon.—This form of diagram is illustrated in Figure 52-B, using the same data as in 52-A. The histogram has been lightly blocked in as a background to show that the polygon can be drawn by connecting the midpoints of the successive columns. It could equally well be drawn without the histogram, by plotting points measured from the abscissas at the midpoints of the class intervals and the ordinates of the corresponding frequencies. The line connecting these points is extended to zero at the midpoint of the class at either end beyond the range of frequencies; thus the broken line or "curve," together with the base line, forms a "polygon" inclosing the entire area of frequencies. /

FIGURE 52

TWO TYPES OF FREQUENCY DIAGRAM OF RENT DATA



Data from Table 60-C.

It can be demonstrated that the total area of the polygon is exactly equal to that of the histogram, although the area included in each class has been slightly altered. When the midpoint of the rectangle standing on the base, \$17.50 to \$27.50, is joined with the midpoint of the rectangle standing on the base, \$27.50 to \$37.50, the triangle CDH is added to the area on the \$17.50 to \$27.50 base and the triangle ABH is removed from the area on the base \$27.50 to \$37.50. But the areas of the triangles are equivalent ($AB = CD$ and angle $a =$ angle b), therefore the area of the figure X_2CBX_4 is equal to the area of the figure X_2CDABX_4 . A similar argument for each adjacent pair of rectangles proves that the area between the polygon and the base

line is the same as the sum of the areas of the rectangles. Hence the polygon is a smooth trace of the histogram in which the total area is preserved but the idea of graduated increase and decrease in the frequencies is substituted for the steps of the histogram.

It must be noted, however, that (unless by chance points $G-C-B$ lie in a straight line) triangle CDH which has been added to the class \$17.50 to \$27.50 is not equivalent to triangle CEK which has been removed from it, and that area X_1KCHX_3 is therefore not equal to area X_1EDX_3 .

This fact leads to the conclusion that the histogram is the more appropriate form to use when it is necessary to represent exactly the number of frequencies in each class, whereas the polygon gives a better picture when a smoothed distribution is wanted.

Uses of Each Type of Graph

For the majority of frequency distributions of economic or social data, that is, for collected data that are in reality samples taken from a larger universe, either type of diagram may be employed. However, because the polygon smooths the contour of a distribution while maintaining the total area, it is suitable only for data that are continuous.

Continuous Data.—In the rent distribution suppose that instead of 155 items the sample had been doubled giving a total of 310 rentals, its representative character, of course, being preserved. If the number of class intervals were then doubled and the width of each interval reduced to five dollars,⁶ the resulting histogram would retain the general shape of Figure 52-A, but due to the operation of the principle of statistical regularity some of the variability of sampling would be removed. Consequently the new histogram would resemble the polygon in Figure 52-B more closely than does the original histogram. If the cases in any such distribution could be multiplied indefinitely, and the class widths decreased accordingly, the final contour would be practically identical whether drawn as a histogram or a polygon. This illustrates the assumption underlying the drawing of a polygon—it interpolates from the sample data the probable intervening values of the universe. Thus it gives a description of the universe derived from the information supplied by a single sample of continuous data. /

⁶ Further reduction of the width of the interval would not be possible in this distribution, regardless of the size of the sample, because of the concentration of the actual amounts on the five-dollar values, \$35, \$40, etc.

This does not mean, however, that each point of the polygon can be read as an actual frequency unless the total frequency is infinite and the width of each class interval infinitesimal. For example, in the rent polygon there might be a natural tendency to think of every point on the polygon as representing a number of frequencies, a tendency to think of 35 rentals, for instance, at \$27.50. This is erroneous because in the polygon as in the histogram there are only 35.5 cases between \$22.50 and \$32.50. It is correct to say that at the \$27.50 level rentals occurred, at the rate of 35 per \$10 interval.

TABLE 62
NUMBER OF EMPLOYEES IN NEW JERSEY LAUNDRIES, NOVEMBER, 1936*

No. of EMPLOYEES		No. of LAUNDRIES
1 and less than 10.....	10.....	16
10 and less than 25.....	25.....	41
25 and less than 50.....	50.....	24
50 and less than 100.....	100.....	7
100 and less than 200.....	200.....	4
200 and less than 300.....	300.....	4
300 and less than 500.....	500.....	2
Total		98

* *Monthly Labor Review*, United States Department of Labor (October, 1937), p. 888.

Discrete Data.—In dealing with discrete data, a polygon would incorrectly indicate intervening values that could not possibly exist; consequently the histogram must be used. Each class of a discrete distribution may include only one unit, or several discrete units grouped together.]

An example of the latter is shown in Table 62, number of laundries employing a specified number of workers. As grouped in this table, each class contains several distinct sizes of laundry. Therefore the columns of a histogram which indicate that the frequencies in each group are distributed over the entire class would give a correct representation. However, these data might be further broken down into the number of laundries employing 8, 9, 10, etc., workers, each class having a width of only one natural unit (the individual worker). In this case the discrete nature of the data would be more accurately represented by separate bars erected at the midpoint of each class interval. The frequencies are all concentrated at these points, and not distributed from 8.5 to 9.5, 9.5 to 10.5, etc.

It should be noted that this is the only situation in which separated bars may be used in a frequency diagram. The adjacent columns of

a **histogram** do resemble bars in a general way, but are of an entirely different nature. The *area* of the columns of a histogram is important, whereas in a bar diagram only the height is measured. When the class intervals are equal, as in all the illustrations up to this point, the heights of the columns are in the same proportion to one another as their areas, because all of their bases are equal. In the case of unequal class intervals, however, the bases of the columns are unequal, hence the areas are not in direct proportion to their heights. This point will be explained later in the chapter.

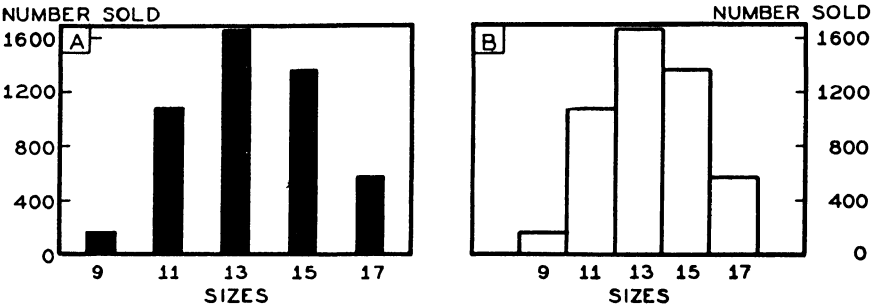
TABLE 63
NUMBER OF JUNIOR DRESSES SOLD DURING MONTH OF FEBRUARY, ACCORDING TO SIZE *

SIZES	NUMBER OF DRESSES SOLD
9	171
11	1,082
13	1,676
15	1,355
17	584
Total	4,868

* Confidential information from a Buffalo department store.

Another example of a discrete natural unit is shown in Table 63. Sizes of junior dresses are discrete classes, hence the use of separated bars is warranted, as shown in Figure 53-A. However, in order to give the impression of area representing the total number of dresses sold, the use of the histogram, Figure 53-B is more common. In this case, the odd-numbered size is the natural unit and is just as indivisible as was the individual employee in the preceding example. There are no

FIGURE 53
FREQUENCY DIAGRAMS OF DISCRETE DATA: NUMBER OF DRESSES SOLD IN JUNIOR SIZES



Data from Table 63.

junior dress sizes other than those given in Table 63⁷ so that the number of classes in such a distribution could not be increased nor the class width decreased, no matter how many cases might occur in the sample.

Any business which sells size merchandise such as men's shirts, gloves, shoes, etc., can utilize this type of analysis in controlling its purchases and inventory. From his past records of regular sales (not including year-end or clearance sales) a merchant can prepare frequency distributions and histograms of sizes of merchandise as guides to his next year's purchases or to the maintenance of his regular stocks. The same distributions and graphs would not be useful in other stores or for other neighborhoods, for the distribution of sizes sold by a particular merchant is peculiar to the characteristics of his clientèle. The distributions will be different, depending upon age, nationality, economic status, occupations, and other characteristics of the people who purchase in each neighborhood or at specific stores.

Adaptations of Frequency Graphs

The Cumulative Graph—Ogive.—For some kinds of analysis and description the cumulative frequency distribution and curve (usually called ogive)⁸ are of more value than the forms just described. A cumulative frequency distribution can be constructed from an ordinary frequency distribution by adding the frequencies of successive class intervals, beginning at the smallest (the largest) class of the distribution, and showing each of these successive totals as the number of cases which is smaller than (greater than) the value of the upper (lower) class limit at that point. Table 64, which contains two types of cumulative distributions constructed from Table 61, shows how they are derived. The frequencies are cumulated from the lower limit to the upper limit of the table in column 2, and from the upper limit to the lower limit in column 3. Table 65 is the presentation form for the results demonstrated in Table 64. The information obtainable from Table 65 is in more usable form than that provided by Table 61. For instance, without any kind of arithmetic treatment, it is immediately obvious from column 1 that more than half of the families in the sample (87) paid monthly rentals of less than \$37.50, and from

⁷ The even numbered sizes are used in a different classification, misses' dresses.

⁸ The name ogive is an architectural term given to the rib of a pointed vault or gothic arch, which has the same shape as this type of curve.

TABLE 64
CUMULATIVE FREQUENCY DISTRIBUTIONS OF MONTHLY RENTALS PAID
BY 155 FAMILIES IN COLUMBUS, OHIO

CLASS INTERVAL	(1) FREQUENCY	(2) CUMULATIVE FREQUENCY, LESS THAN UPPER LIMIT	(3) CUMULATIVE FREQUENCY, LOWER LIMIT AND ABOVE
\$ 7.50 and under \$17.50.....	16	16	155
17.50 and under 27.50.....	27	43	139
27.50 and under 37.50.....	44	87	112
37.50 and under 47.50.....	17	104	68
47.50 and under 57.50.....	18	122	51
57.50 and under 67.50.....	11	133	33
67.50 and under 77.50.....	10	143	22
77.50 and under 87.50.....	9	152	12
87.50 and under 97.50.....	3	155	3
Total	155		

column 3 that about one-fifth of the group (33) paid rentals of \$57.50 per month or over.

The curves of these two types of cumulative frequency distributions are shown in Figure 54. Curve *L* represents the "less than" distribution and curve *M* the "more than" distribution. It should be noted that in Figure 52-B the *midpoints* of the class intervals were joined to form the frequency polygon, whereas in the ogives the *end* values are joined. In the "less than" ogive there are, for example, sixteen cases below \$17.50; therefore the frequency 16 is plotted at the upper limit of the \$7.50 to \$17.50 class. Similarly the next frequency, 43, is plotted at \$27.50, etc.

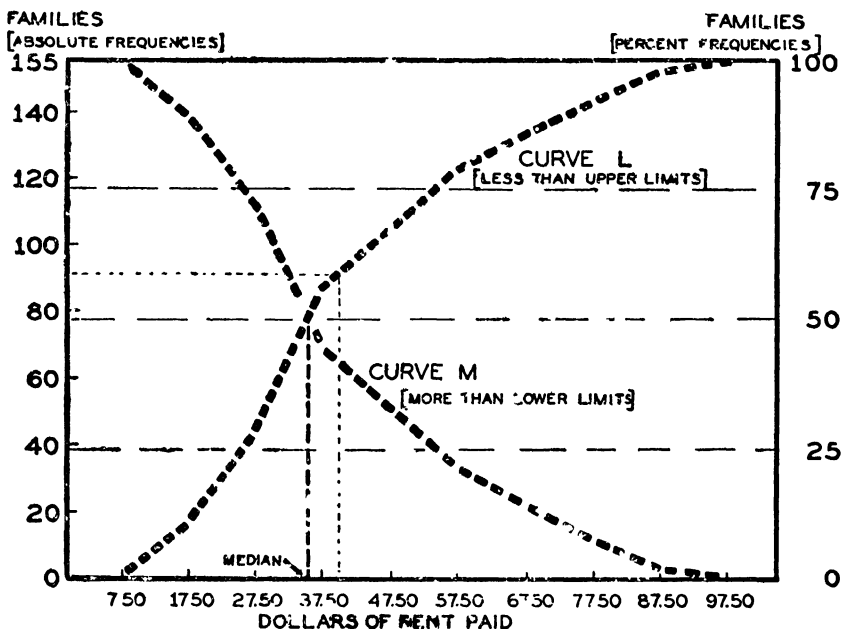
There are two major characteristics of the ogive, the most important of which is the ease of interpolation which its use permits. In order

TABLE 65
NUMBER AND PER CENT OF FAMILIES IN COLUMBUS, OHIO, PAYING
MORE THAN AND LESS THAN A SPECIFIED MONTHLY RENTAL

RENTALS PAID	FAMILIES		RENTALS PAID	FAMILIES	
	(1) Number	(2) Per Cent		(3) Number	(4) Per Cent
Less than \$17.50....	16	10.3	\$ 7.50 and more....	155	100.
Less than 27.50....	43	27.7	17.50 and more....	139	89.7
Less than 37.50....	87	56.1	27.50 and more....	112	72.3
Less than 47.50....	104	67.1	37.50 and more....	68	43.9
Less than 57.50....	122	78.7	47.50 and more....	51	32.9
Less than 67.50....	133	85.8	57.50 and more....	33	21.3
Less than 77.50....	143	92.3	67.50 and more....	22	14.2
Less than 87.50....	152	98.1	77.50 and more....	12	7.7
Less than 97.50....	155	100.	87.50 and more....	3	1.9

FIGURE 54

OGIVES: CUMULATIVE FREQUENCY DIAGRAM OF RENT DATA



Data from Table 65.

to determine the number of cases in which less than a given amount, say \$40, is paid for rent, it is only necessary to make a vertical ruling from \$40 on the horizontal scale to ogive L and then from this point of intersection make a horizontal ruling to the Y axis. This indicates a frequency of approximately 91 families who pay less than \$40 per month. The second important characteristic is the slope of the ogive. Where the slope is steepest there is the greatest concentration of frequencies, and wherever it is less steep there are fewer frequencies.

The cumulative distribution and the ogive are often presented on a percentage basis in practical work. To illustrate this usage, columns 2 and 4 have been included in Table 65. In Figure 54 the scale at the right has been so arranged that 100 per cent is in the same position as 155 on the left-hand scale. The ogives L and M represent respectively either columns 1 and 3 or columns 2 and 4 of Table 65. Information concerning the distribution of rents in the sample can be obtained by reading the left scale as previously indicated. The percentage scale is independent of the actual number of cases in the

sample. From it can be read facts concerning the distribution of rents in Columbus, on the assumption that the sample is representative of the entire city. Thus ogive *L* shows that 25 per cent of Columbus families pay monthly rents of not more than \$26.50 and ogive *M* shows that 25 per cent pay at least \$54.00. The point of intersection of the two ogives at a frequency of 50 per cent shows that half of the families pay more and half less than about \$35.25 per month.

The ogive has an additional use in the graphic determination of measures of central tendency and dispersion. In particular this diagram will be referred to in chapter XVII.

Histogram of Unequal Classes.—In the discussion of the principles for constructing frequency distributions provision was made for unequal class intervals under certain conditions. The graph of such a distribution requires special explanation because the use of the methods previously described would produce definite misrepresentation.

TABLE 66

HOURLY WAGE RATES PAID TO NEWLY HIRED EMPLOYEES BY 52 INDUSTRIAL CONCERNS IN BUFFALO, NEW YORK, IN 1940*

DISTRIBUTION A EQUAL INTERVALS		DISTRIBUTION B UNEQUAL INTERVALS		
Hourly Wage Rate in Cents	(1) No. of Concerns	Hourly Wage Rate in Cents	(2) No. of Concerns	(3) Heights of Columns Adjusted to Preserve Frequency Area in Unequal Intervals
27.5–32.4	2	27.5–37.4	4	2
32.5–37.4	2	37.5–47.4	12	6
37.5–42.4	2	47.5–48.4	3	15
42.5–47.4	10	48.5–49.4	5	25
47.5–52.4	24	49.5–50.4	8	40
52.5–57.4	8	50.5–51.4	6	30
57.5–62.4	2	51.5–52.4	2	10
62.5–67.4	0	52.5–62.4	10	5
67.5–72.4	2	62.5–72.4	2	1
Total	52		52	

* Source confidential.

A tabulation of hourly hiring rates in Buffalo is presented in Table 66. Distribution A is divided into equal five-cent class intervals. Nearly half of the cases fall in one class and as a result the table does not provide as much information as we should like to have from a frequency table concerning hiring rates. The effect of this concentration is even more evident in the histogram of Figure 55-A. The middle rectangle is so large in relation to the others that the

comparison of frequencies by means of the areas of the several rectangles discloses little information beyond the fact, evident from the table, that the hiring rates are concentrated around 50 cents an hour.

In order to learn more about hiring rates, Distribution B was prepared from the original source. Ten-cent intervals were used below 47.5 cents and above 52.5 cents, but the middle class was subdivided into 5 one-cent intervals to provide additional detail concerning the 24 cases in that class. That is, unequal class intervals were introduced as a means of increasing the usefulness of the table.

This distribution is plotted in Figure 55-B in the usual way, i.e., with the heights of the rectangles of the histogram representing the frequencies just as written in column 2 of the table. The appearance of the graph is sufficient evidence of its inaccuracy. The difficulty lies in the fact that the areas allotted to the frequencies in the several classes do not correspond to the equivalent areas in Figure 55-A. The contrast can be followed in a tabulation of the two graphs (Table 67).

The areas within the several classes are not the same for A and B nor are the two total areas. The last two columns indicate, however,

TABLE 67
AREAS OF CORRESPONDING RECTANGLES OF FIGURE 55
(Frequency $\times$ width of class interval)

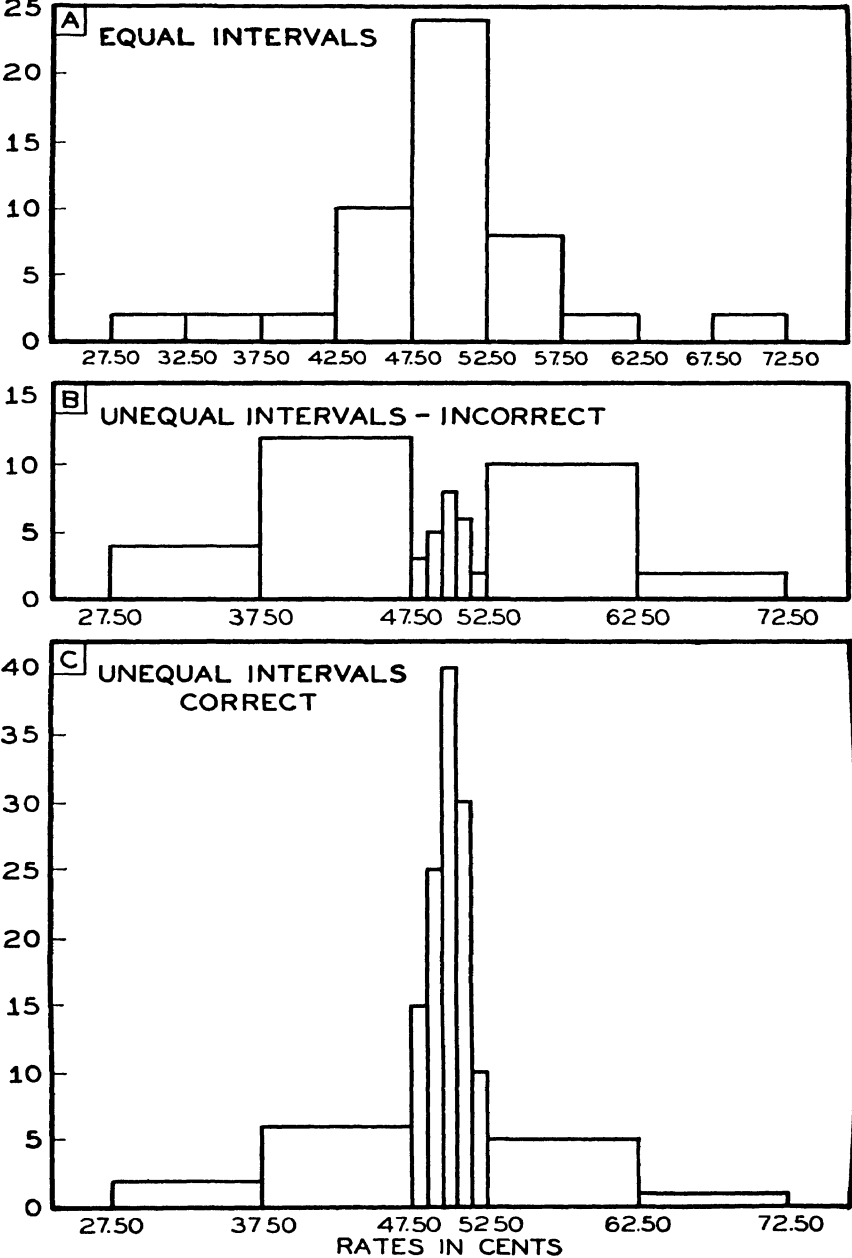
CLASS INTERVALS	(1) FIGURE 55-A	(2) FIGURE 55-B	(3) FIGURE 55-C
27.5-37.4	$\left\{ \begin{array}{l} 2 \times 5 = 10 \\ 2 \times 5 = 10 \end{array} \right\} = 20$	$4 \times 10 = 40$	$2 \times 10 = 20$
37.5-47.4	$\left\{ \begin{array}{l} 2 \times 5 = 10 \\ 10 \times 5 = 50 \end{array} \right\} = 60$	$12 \times 10 = 120$	$6 \times 10 = 60$
47.5-52.4	$24 \times 5 = 120$	$\left\{ \begin{array}{l} 3 \times 1 = 3 \\ 5 \times 1 = 5 \\ 8 \times 1 = 8 \\ 6 \times 1 = 6 \\ 2 \times 1 = 2 \end{array} \right\} = 24$	$\left\{ \begin{array}{l} 15 \times 1 = 15 \\ 25 \times 1 = 25 \\ 40 \times 1 = 40 \\ 30 \times 1 = 30 \\ 10 \times 1 = 10 \end{array} \right\} = 120$
52.5-62.4	$\left\{ \begin{array}{l} 8 \times 5 = 40 \\ 2 \times 5 = 10 \end{array} \right\} = 50$	$10 \times 10 = 100$	$5 \times 10 = 50$
62.5-72.4	$\left\{ \begin{array}{l} 0 \times 5 = 0 \\ 2 \times 5 = 10 \end{array} \right\} = 10$	$2 \times 10 = 20$	$1 \times 10 = 10$
Total Area.....	260	304	260

exactly what should be done to Distribution B in order to obtain a graph whose area will be comparable with the graph of Distribution A. The areas of rectangles on class intervals that have been increased from five cents to ten cents in Figure 55-B have been doubled, and the areas

FIGURE 55

FREQUENCY DIAGRAMS OF HOURLY WAGE RATES PAID BY FIFTY-TWO INDUSTRIAL CONCERNS

NUMBER OF CONCERNS



Data from Tables 66 and 67.

of rectangles on the middle class interval which has been subdivided into 5 one-cent intervals total one-fifth of the former amount. Therefore, dividing the frequencies of the first, second, eighth, and ninth classes by 2, (i.e., multiplying by $\frac{1}{2}$), and multiplying each of the intervening classes by 5 (i.e., by $\frac{5}{1}$), gives the set of heights of columns adjusted to compare with frequencies of even five-cent intervals. These adjusted heights, as listed in column 3 of Table 66, are used for the histogram shown in Figure 55-C. The areas of this diagram agree with those in Figure 55-A (shown in Table 67). The first two rectangles of A are identical with the first rectangle of C. The areas of the third and fourth rectangles of A are equivalent to the area of the second rectangle of C. Similar computations show the equivalence of all the corresponding rectangles in the two diagrams. Thus in C the areas of the total and the individual classes respectively have been preserved. At the same time additional information has been presented concerning the large number of hiring rates in the middle class without sacrificing any essential facts relative to the rates in the other class intervals.

The general rule for adjusting the frequencies for use in a diagram⁹ of a distribution containing unequal class intervals may be stated as follows: divide the actual frequency of each class by the width of the class to obtain unit frequencies; multiply these unit frequencies by the width of the equal class intervals of another distribution with which they are to be compared. If no comparison is involved the unit frequencies themselves may be plotted or any constant multiple of them.

Comparison of Two Distributions.—In applied statistical work as well as in more advanced analysis occasions arise which require that two distributions be represented on the same graph. The purpose is to show the relation between the contours of the two curves and the positions of measures descriptive of the two distributions. Polygons should be used for comparison because the rectangles of histograms would overlap making clear-cut representation impossible.

In comparing two polygons certain requirements must be met. The class intervals of the two distributions must be the same, and all of the intervals must have the same width. Percentage frequencies must be used to give comparable areas. Two distributions that are to be

⁹ When the class intervals are unequal the transfer from the histogram to the polygon is not justified because in joining the midpoints of adjacent columns the triangular areas added and subtracted are not equivalent.

compared must be brought into conformity with these rules before the graph is planned.

TABLE 68

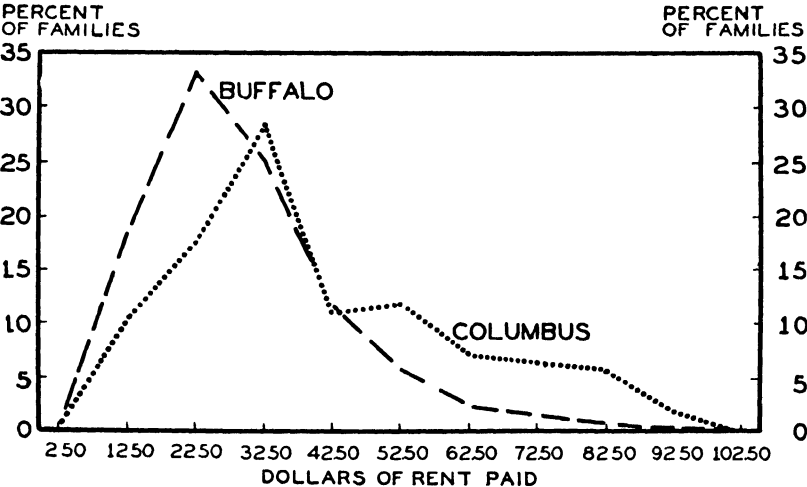
NUMBER AND PERCENTAGE DISTRIBUTION OF 500 FAMILIES IN BUFFALO, NEW YORK, ACCORDING TO VALUE OF MONTHLY RENTALS PAID (Approximated from reports of real estate dealers)

MONTHLY RENT	FAMILIES	
	Number	Percentage Distribution
\$ 7.50 and less than \$17.50.....	93	18.6
17.50 and less than 27.50.....	167	33.4
27.50 and less than 37.50.....	127	25.4
37.50 and less than 47.50.....	59	11.8
47.50 and less than 57.50.....	30	6.0
57.50 and less than 67.50.....	12	2.4
67.50 and less than 77.50.....	7	1.4
77.50 and less than 87.50.....	4	.8
87.50 and less than 97.50.....	1	.2
Total	500	100.

Figure 56 contains two polygons representing comparable material. One is a reproduction of Figure 52-B in per cents, showing the rentals paid in Columbus, and the other is a diagram of the data in Table 68, showing the distribution of rentals in Buffalo. The Buffalo distribution is only approximate because the reports of real estate offices may contain some duplication. In the main, however, the sample is a representative cross-section of rents in Buffalo.

FIGURE 56

PER CENT COMPARISON OF TWO DISTRIBUTIONS OF RENT DATA
Data from Tables 61 and 68.



Comparison of the two polygons shows that the general level of rents is lower in Buffalo than in Columbus. The Buffalo curve is somewhat smoother since it is based on a larger sample. Regardless of the difference in size of sample there is a noticeably greater concentration of rents in Buffalo in the classes below \$47.50, and correspondingly a much smaller proportion in the higher rent brackets.

The higher rent level in Columbus is presumably an expression of an increase in demand for housing due to expanding personnel of the state government unaccompanied by an equivalent growth in housing facilities. If this explanation is correct, the difference in conditions is presumably temporary. If other basic causes are present, further investigation might uncover a permanent differential in the rent levels of the two cities.

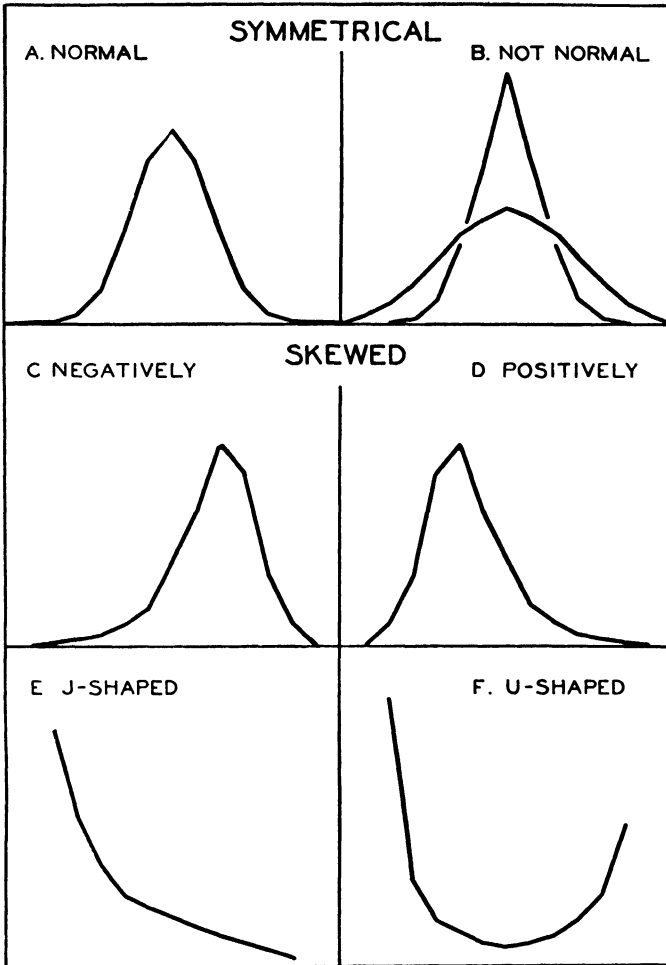
Types of Curves

One branch of advanced statistics deals solely with the various types of frequency curves and the development of measures used in analyzing them. The subject is introduced here in order to acquaint the student with the graphic appearance of the types of curves most frequently encountered in practical work.

Any description of the different types of curves centers around the "normal" or "bell-shaped" curve. It is a portrayal of the distribution of an infinite number of identically obtained measurements of a fixed object. That is, if an extremely large number of persons, all equally skillful, all possessed of normal vision and all using the same minutely graduated measuring device, were to measure the width of a room, the results would vary above and below the true width of the room as indicated by the normal curve. The curve is, therefore, a picture of the variations due to pure chance. But in practice pure chance is usually mingled with other uncontrolled causes of variation to such an extent that a normal distribution is seldom found outside the laboratory. Yet many of the distributions with which we deal differ so little from the normal curve that its characteristics are transferable, and in addition the normal curve serves as a guide to the description of other types of curves.

In Figure 57 six other curves are presented with the normal curve. The two curves in B are symmetrical like the normal curve but one is flatter and the other more sharply peaked. The flat-topped curve would result from a distribution in which extraneous factors

FIGURE 57
TYPES OF CURVES



had produced more variability than would arise from pure chance. The peaked curve would result from a distribution in which extraneous factors tended to offset natural chance variability. C and D are skewed curves depicting distributions in which a controlling factor enters more strongly on one side of the peak than on the other side. These curves have a "long side" and a "short side." Methods of interpreting these are explained under the subject of skewness in chapter XVIII. E and F are less usual but are types that are encountered occasionally.

Detailed analysis of these curves belongs in more advanced statistical work, but some of the properties of the normal curve and its use in the development of the principles of reliability of samples will be the subject matter of chapters XXVIII and XXIX. Ability to recognize the several types is essential to an understanding of the chapters dealing with averages and dispersion.

The Lorenz Curve

A special type of diagram used to show the nature of the concentration of cases in one or more frequency distributions is known as the Lorenz Curve.¹⁰ The method of preparing data for presentation in a Lorenz Curve can be explained best from an example. Table 69 gives the number of independent retail grocery stores operating in Buffalo in 1929 and 1935 classified according to size as measured by sales.

Column 3 is obtained by changing column 2 to per cents and in column 4 these per cents are cumulated. Column 7 is the result of three steps: (a) multiply the midpoint, column 1, of each class by the number of stores, column 2, in the class to obtain the total sales by stores of that size, column 5; (b) express each of these products as a percentage of the total sales of all stores, column 6; (c) cumulate these per cents. A parallel procedure using the frequencies for 1935 leads to the cumulative per cents of the lower part of the table.

In plotting the points in Figure 58, the cumulative per cents of stores, column 4, are located from the base scale and the cumulative per cents of sales, column 7, from the vertical scale. Each curve therefore begins at the lower left-hand corner of the diagram and ends at the upper right-hand corner.¹¹ If all of the stores had equal sales, then any 10 per cent of the stores would have 10 per cent of the sales volume, any 20 per cent would have 20 per cent of the sales volume and so on, and the plotted points would fall on the diagonal line of the diagram. Hence this diagonal is known as the line of equal distribution. The departure of the actual curves from this line shows the extent of the concentration of sales volume in the larger-sized

¹⁰ This curve is named after M. O. Lorenz, who developed it and employed it mainly in his studies of wealth. See M. O. Lorenz, "Methods of Measuring the Concentration of Wealth," *Journal of the American Statistical Association*, New Series No. 70 (June, 1905), pp. 209-19.

¹¹ The base scale is sometimes arranged in reverse order, from right to left, so that the curves will extend from the right of the base scale to the top of the vertical scale.

TABLE 69
COMPUTATION OF DATA FOR LORENZ CURVES: SALES AND NUMBER OF STORES BY SIZE GROUPS AS MEASURED BY ANNUAL SALES, FOR
INDEPENDENT RETAIL GROCERY STORES IN BUFFALO, N. Y., 1929 and 1935 *

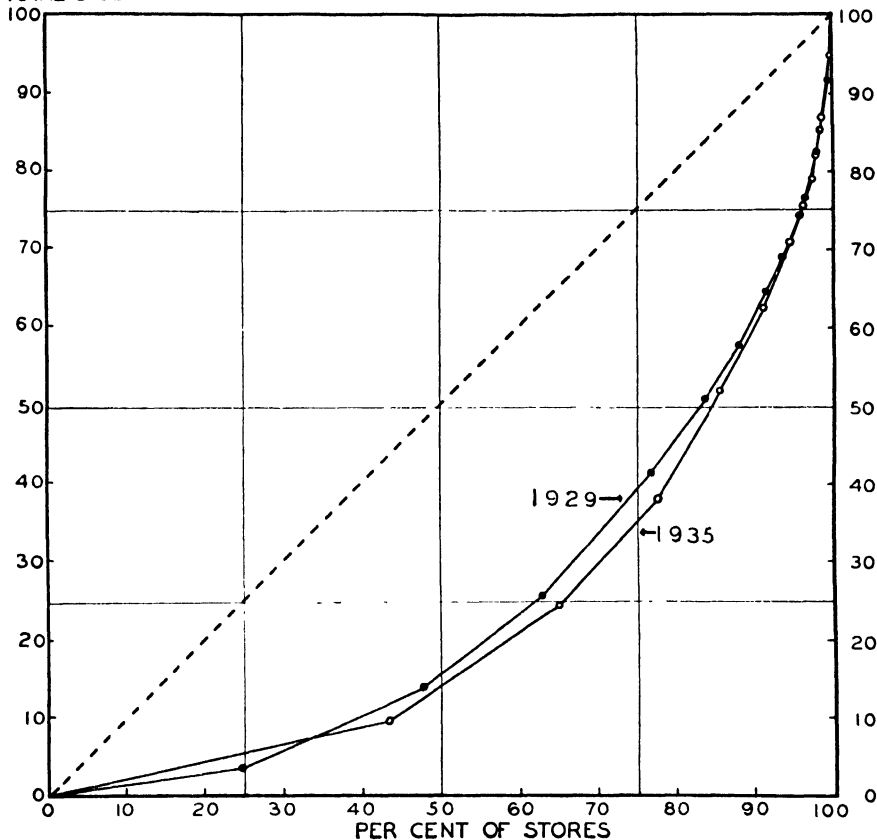
ANNUAL SALES	(1) MIDPOINT	(2) NUMBER OF STORES	(3) PERCENTAGE DISTRIBUTION OF STORES FROM (2)	(4) CUMULATIVE PER CENTS OF STORES FROM (3)	(5) DOLLAR VALUE OF SALES (1) X (2)	(6) PERCENTAGE DISTRIBUTION OF SALES FROM (5)	(7) CUMULATIVE PER CENTS OF SALES FROM (6)
1929							
Less than \$ 5,000.....	\$ 2,500	396	24.6	24.6	\$ 990,000	3.8	3.8
5,000 and less than 10,000.....	7,500	363	22.6	47.2	2,722,500	10.4	14.2
10,000 and less than 15,000.....	12,500	248	15.4	62.6	3,100,000	11.8	26.0
15,000 and less than 20,000.....	17,500	228	14.2	76.8	3,990,000	15.2	41.2
20,000 and less than 25,000.....	22,500	113	7.0	83.8	2,542,500	9.7	50.9
25,000 and less than 30,000.....	27,500	68	4.2	88.0	1,870,000	7.1	58.1
30,000 and less than 35,000.....	32,500	52	3.2	91.2	1,690,000	6.5	64.5
35,000 and less than 40,000.....	37,500	38	2.4	93.6	1,425,000	5.4	70.0
40,000 and less than 45,000.....	42,500	27	1.7	95.3	1,147,500	4.4	74.3
45,000 and less than 50,000.....	47,500	10	0.6	95.9	475,000	1.8	76.2
50,000 and less than 60,000.....	55,000	29	1.8	97.7	1,595,000	6.1	82.2
60,000 and less than 100,000.....	80,000	26	1.6	99.3	2,080,000	7.9	90.2
100,000 and over.....	234,132†	11	0.7	100.	2,575,500	9.8	100.
Total		1,609	100.		\$26,203,000	100.	
1935							
Less than \$ 5,000.....	\$ 2,500	715	43.0	43.0	\$ 1,787,500	9.7	9.7
5,000 and less than 10,000.....	7,500	365	21.9	64.9	2,737,500	14.9	24.6
10,000 and less than 15,000.....	12,500	204	12.3	77.2	2,550,000	13.9	38.5
15,000 and less than 20,000.....	17,500	137	8.2	85.4	2,397,500	13.0	51.5
20,000 and less than 25,000.....	22,500	91	5.5	90.9	2,047,500	11.1	62.6
25,000 and less than 30,000.....	27,500	56	3.4	94.3	1,540,000	8.4	71.0
30,000 and less than 35,000.....	32,500	25	1.5	95.8	812,500	4.4	75.4
35,000 and less than 40,000.....	37,500	18	1.1	96.9	675,000	3.7	79.1
40,000 and less than 45,000.....	42,500	12	0.7	97.6	510,000	2.8	81.9
45,000 and less than 50,000.....	47,500	13	0.8	98.4	617,500	3.4	85.2
50,000 and less than 60,000.....	55,000	4	0.2	98.6	220,000	1.2	86.4
60,000 and less than 100,000.....	80,000	19	1.2	99.8	1,520,000	8.3	94.7
100,000 and over.....	196,062†	5	0.3	100.	980,310	5.3	100.
Total		1,664	100.		\$18,395,310	100.	

*Edmund M. McGarry, "The Structure and Stability of Retail Trade in Buffalo, 1929, 1933 and 1935—Grocery Stores" Supplement of the *Statistical Survey*, Buffalo. The University of Buffalo Bureau of Business and Social Research, March, 1939.
† Actual Average.

FIGURE 38

LORENZ CURVES: CUMULATIVE PER CENTS OF STORES AND SALES, INDEPENDENT RETAIL GROCERY STORES IN BUFFALO, 1929 AND 1935

PER CENT OF
TOTAL SALES



Data from Table 69.

stores. The greater distance of the 1935 curve from the diagonal line shows the growth of concentration between 1929 and 1935.

The Lorenz Curve is valuable, both in analysis and in presentation whenever distribution according to two quantitative attributes is of importance. There has been frequent use of this graph during recent years when distribution of business and income have been under discussion.

PROBLEMS

1. a) What are the advantages of Table 57, page 352, of the text as compared with Table 56, page 351?
- b) Describe exactly how you would obtain Table 57 from Table 56.

- 2. What information concerning rentals can be obtained from Figure 49, page 353? from Figure 50, page 354?
- 3.
 - a) Name the four principles that must be observed in planning a frequency distribution.
 - b) State the main points to be considered in applying each principle.
- 4. Indicate which of the following are correct statements and amend any that are incorrect:
 - a) The presence of artificial grouping in an array can be disregarded in preparing a frequency distribution.
 - b) All frequency distributions should have at least ten class intervals.
 - c) Class intervals can be of equal or unequal width at the convenience of the person preparing a distribution.
 - d) Class limits should be established so that the average value of the items included in each interval is approximately equal to the class mark of the interval.
 - e) In preparing a distribution of continuous data the only way to designate class limits is by writing the class marks.
 - f) The following is a discrete distribution:

POPULATION	No. OF CITIES
Over 1,000,000	5
500,000 up to 1,000,000.....	9
300,000 up to 500,000.....	16
etc.	

- 5. State wherein each of the following meets or fails to meet the principles of construction of a frequency distribution.

A	
INCOME	AVERAGE MONTHLY RENT
Under \$ 500..	\$25.90
\$ 500 to 700..	22.90
700 to 1,000..	22.80
1,000 to 1,200..	26.00
1,250 to 1,500..	28.10
etc.	

B	
AGE (YEARS)	No. OF PERSONS
All ages	6,930,446
Under 5	535,600
Under 1	100,398
5 to 9	577,284
10 to 14	575,300
15 to 24	1,287,625
25 to 34	1,345,984
etc.	

C	
TYPE OF DWELLING	No. OF FAMILIES PROVIDED FOR
One-family	4,620
Two-family	159
Multi-family	1,195
Total	5,974

D		
YEAR OF GRADUATION	EARNING OVER \$4,000	
	Per Cent of Class	Per Cent of Group
1935	2	18
1936	5	26
1937	4	29
1938	2	16
1939	1	11
Total		100

6. a) Describe the construction of a histogram; a frequency polygon.
b) What is the relation of the two types of diagram to discrete and continuous data?
7. For what kinds of information is the ogive preferable to the ordinary distribution?
8. a) Under what circumstances is it desirable to use unequal class intervals?
b) Explain with an example of your own the method of preserving areas in a diagram of a frequency distribution with unequal class intervals.
9. a) What is the reason for using percentage frequencies in comparing two distributions?
b) In what situation would percentage frequencies be unnecessary for comparing two distributions?
10. a) Make a frequency table, using the 112 items in the 4 columns assigned to you from the following table. (See numbered assignments at the top of page 386.)
b) Give reasons for your choice of class limits and width of class intervals.
c) Draw a graph showing your frequency distribution.
d) What information concerning wages of semi-skilled female workers in this hosiery mill can be derived from your table and graph?

WEEKLY EARNINGS OF 168 SEMI-SKILLED FEMALE WORKERS, IN HOSEYRY MILL XYZ *

(a)	(b)	(c)	(d)	(e)	(f)
15.20	18.00	11.20	16.00	20.00	13.60
11.60	14.00	12.00	11.30	12.20	12.00
8.00	12.00	17.60	15.60	8.50	8.00
12.80	12.80	9.50	12.00	14.50	10.00
14.00	11.80	12.00	10.60	16.00	12.60
6.40	9.20	14.00	12.00	12.60	14.00
12.00	7.60	12.00	15.00	12.00	6.50
12.40	14.80	8.20	6.00	8.00	16.00
24.00	18.00	28.00	8.00	19.00	14.00
14.60	16.80	16.80	16.00	22.00	14.60
9.00	14.20	14.40	17.20	15.20	19.20
16.50	12.00	21.20	14.40	10.00	12.30
20.00	12.00	20.00	12.50	14.00	11.60
18.00	21.00	23.00	20.00	16.00	16.40
14.10	8.00	14.00	18.80	16.40	16.00
22.50	16.00	16.10	12.00	12.00	20.00
12.00	24.00	19.90	12.00	23.80	21.40
20.80	19.60	12.90	8.40	28.40	24.00
16.00	27.00	24.00	23.50	17.30	28.80
18.00	20.00	16.00	20.00	18.00	15.20
7.20	10.40	8.00	21.60	14.00	25.00
14.00	15.50	11.80	24.40	11.40	12.00
26.00	21.80	15.00	14.00	24.50	20.40
16.00	14.00	16.00	16.20	6.00	17.60
16.00	6.00	12.40	28.00	20.00	8.80
12.00	16.00	18.40	16.90	16.00	16.00
19.40	12.40	15.50	13.00	12.00	18.00
10.00	16.00	6.00	14.00	13.20	12.00

* Based on similar data appearing in a 1939 issue of the *Monthly Labor Review*.

Assignments

No. Columns		No. Columns		No. Columns	
1	a b c d	6	a b e f	11	b c d e
2	a b c e	7	a c d e	12	b c d f
3	a b c f	8	a c d f	13	b c e f
4	a b d e	9	a c e f	14	b d e f
5	a b d f	10	a d e f	15	c d e f

REFERENCES

CHADDOCK, ROBERT E., *Principles and Methods of Statistics*. Boston: Houghton Mifflin Co., 1925.

Chapter V deserves a careful reading, particularly the location of partition values, pp. 61–65.

ELDERTON, W. PALIN, and ELDERTON, ETHEL M., *Primer of Statistics*. London: A. and C. Black, Ltd., 1920.

Chapters I–IV give an extremely simple statement of the fundamentals of the preparation of frequency distributions and the meaning of measures of central tendency and dispersion.

MUDGETT, BRUCE D., *Statistical Tables and Graphs*. Boston: Houghton Mifflin Co., 1930.

The explanation of "Graphs of Frequency Distributions" on pp. 102–21 warrants careful reading.

TRELOAR, ALAN E., *Elements of Statistical Reasoning*. New York: John Wiley and Sons, 1939.

A clear explanation of the difference between discrete and continuous distributions and adjustment for intervals of different width, pp. 26–35.

YULE, G. UDNY, and KENDALL, M. G., *Introduction to the Theory of Statistics*. London: Charles Griffin and Company, Ltd., 1937.

The basic principles of the construction of frequency distributions and graphs are presented in chapter VI.

CHAPTER XVI

MEASURES OF CENTRAL TENDENCY—AVERAGES OF CALCULATION

INTRODUCTION

THE whole process of statistical analysis is characterized by the attempt to reduce the details of masses of data and to develop summary figures. The initial stages in this analysis have already been pointed out: a statistical table classifies masses of separate items into a small number of comparable groups; a graph is planned to concentrate attention on one or a few major characteristics of a set of data; a ratio involves the substitution of one simple figure for two or more unwieldy ones; and a frequency distribution condenses a long list of separate items into usable form by substituting class values for individual values. The statistician's working equipment must include a knowledge of these various descriptive devices. Another basic tool needed in analysis is the average. An average is frequently described as a "measure of central tendency" because it provides a single summary figure by means of which an entire set of data may be represented.

Measures of central tendency are familiar to statisticians and laymen alike in such examples as average weekly wages, average prices of securities, average daily temperature, a man of average height, a medium-sized house, and the usual amount of rainfall. Familiarity, however, tends to obscure the fact that several different concepts of "average" are involved in these examples and that different methods of computation are employed in obtaining them. It follows, therefore, that several types of summary figures or averages must be explained in developing the subject.

Measures of central tendency fall into two groups: (1) those obtained by calculation, and (2) those defined by position. Each group contains two fundamental averages that have sufficient practical application to warrant explanation in this book. They are,

Averages of Calculation
Arithmetic Average
Geometric Average

Averages of Position
Median
Mode

The remainder of this chapter will be devoted to the arithmetic average and the geometric average. The averages of position and criteria for evaluating the four measures of central tendency will be described in the following chapter.

THE ARITHMETIC AVERAGE

The Average of Ungrouped Data

The measure of central tendency most commonly known and recognized is the arithmetic average, which is frequently called the arithmetic mean or, more simply, the mean. It is calculated by adding together all the items in a group or series, and dividing their sum by the number of items. For instance, the arithmetic average of a student's examination grades is calculated by adding the grades of all the examinations and dividing their sum by the number of examinations. Table 70 illustrates the method.

TABLE 70
CALCULATION OF ARITHMETIC AVERAGE OF EXAMINATION GRADES

First examination	75
Second examination	93
Fourth examination	88
Third examination	87
Fifth examination	93
Total, 5 examinations	436

$Average = 436 \div 5 = 87.2$

There are several characteristics of this simple problem which should be emphasized because they apply to all arithmetic averages. First, all five items are included in the total, which is divided by 5 to obtain the average. Second, a change in any one of the examination grades will affect the value of the average; the average of grades would be increased 5 points by changing the grade of 75 to 100. All students have without doubt made this kind of calculation by assuming different examination grades, before and after an examination. Third, extreme values, either high or low, may produce a value of the average which is not representative of the data. Unusual values have the greatest effect when the average is based on a small number of items.

This method of calculating the arithmetic average can be applied to the sample of rentals paid by families in Columbus as arrayed in Table 57, page 352. The sum of all the 155 monthly rentals is

determined to be \$6,307. This total divided by 155 gives the arithmetic average, \$40.69.¹ If each of the 155 families in the group had paid a monthly rental equal to the arithmetic average, \$40.69, the total amount would be the same as was actually paid. The arithmetic average, then, is the value which can be substituted for every actual value in a group without changing the sum of all the values.

The Weighted Average and Weighted Total

There are many cases in which values to be averaged or totaled are of different degrees of importance. When this is the situation, it is necessary to weight the separate items by multiplying each by a factor which represents its relative importance in the group.

Weighted Average.—The problem of weights usually arises in the calculation of course grades. For example, suppose that the fifth grade, 93, in Table 70, was for a final examination, and consequently was twice as important as any other examination grade received. It would be multiplied by 2 and the total divided by 6 (the sum of the weights) as indicated in Table 71. The weighted average is one grade point larger than the unweighted average. In some cases the weighting might cause a much greater difference than one grade point, and it might cause the average to increase (as in this case) or to decrease.

TABLE 71
CALCULATION OF WEIGHTED AVERAGE OF EXAMINATION GRADES

	EXAMINATION GRADE	WEIGHT	GRADE × WEIGHT
First	75	1	75
Second	93	1	93
Third	87	1	87
Fourth	88	1	88
Final	93	2	186
Total		6	529

$$\text{Weighted arithmetic average} = 529 \div 6 = 88.2.$$

¹ A formula for the arithmetic average can be developed from this calculation,

$$\text{Arithmetic average} = \frac{\Sigma X}{N},$$

where Σ (sigma) stands for "the sum of"; X stands for any value of the variable, rent; and N represents the total number of items (that is, rentals). There are several commonly used symbols for the arithmetic average which are employed under different circumstances. In elementary work it is usually represented by its initials, *A.A.*, or by M (Mean). The latter will be employed in this text. If several averages are being used, a subscript is added to M to designate the variable of which it is the average: e.g., M_x for the arithmetic average of the X 's, M_y for the average of the Y 's, etc. When used in algebraic manipulation the average is sometimes represented by its formula, or by the symbol, $\bar{X}$.

The weighted average of examination grades alone does not provide a complete basis for a course grade. For the latter, it is necessary to include laboratory work, classroom responses, and special assignments or reports, as well as examinations. To average the grades of these diverse elements, each of which is of different importance, it is again necessary to employ weights. Since each type of grade included might be a weighted average, like the examination grade above, the final course grade becomes a weighted average of weighted averages.

A weighted average requires careful consideration of the items being averaged, in order to arrive at an equitable basis for establishing the weights. In the case of a course grade in statistics, after scrutinizing all elements involved, it may have been decided that they should have the measures of importance shown in column 1, Table 72.

The per cents which represent the importance of each course element in the final grade may be called weights. The weighted average is obtained by (1) multiplying the value of each item by its weight, a measure of its importance in the total; (2) dividing the sum of these products by the sum of the weights. The method of calculating a weighted average to obtain a course grade for a student in elementary statistics is shown in Table 72.

TABLE 72
CALCULATION OF WEIGHTED AVERAGE OF ELEMENTS OF COURSE
TO OBTAIN FINAL COURSE GRADE

ELEMENT	(1) WEIGHT (per cent)	(2) AVERAGE GRADE	(3) AVERAGE GRADE × WEIGHT
Examinations	60	88	5,280
Laboratory work	20	80	1,600
Classroom response	10	90	900
Homework	10	70	700
Total	100	...	8,480

$$\text{Weighted average} * = \frac{8,480}{100} = 84.8$$

* If the letter W stands for the weight and n stands for the number of items, the process can be described algebraically as:

$$\text{Weighted average} = \frac{X_1W_1 + X_2W_2 + X_3W_3 + \dots + X_nW_n}{W_1 + W_2 + W_3 + \dots + W_n}$$

$$\text{In summary form: Weighted average} = \frac{\sum(XW)}{\sum W}$$

The weighted average, 84.8, is greater than the simple arithmetic average (the sum of the grades in column 2 divided by 4), by more than 2 points. This increase of 2 points over the simple average is due

to the increased importance which is given to the high average examination grade of 88 through the process of weighting.

The weights of the several parts of the statistics course are shown in Table 72 as per cents, the sum of which is 100, representing the whole course. From elementary arithmetic it will be remembered that the numerator and denominator of a fraction can be multiplied or divided by the same number without changing the value of the fraction. Consequently, the absolute values of the weights can be replaced by relative values proportional to the absolute values. The relatives are easier to use and give the same result.

In a study of wages of farm labor in Vermont for the period 1780–1937, a weighted average is used to calculate the annual average of day wage rates. "The annual wage rates of labor hired by the day are weighted averages of the monthly data."² The weights assigned to the average daily wage rates of the several months are:

January	5	May	8	September	8
February	4	June	10	October	8
March	5	July	18	November	6
April	7	August	16	December	5
					Total
					100

Examination of these weights reveals that they are closely related both to the seasons and to the number of days in each month. The day wages in July, for instance, are of most importance because of the great amounts of farm labor hired in that month, the usual clemency of the weather, and the critical period for growing crops, as well as the number of working days in the month. The day rates in July, then, should have the greatest influence in the determination of the annual average. Conversely, February rates should have the least influence.

Effect of Weights.—One might justifiably ask, "What are the effects of weighting on the simple average?" The answer may be generalized as follows:

1. If the larger weights are applied to the smaller values, and the smaller weights to the larger values, the influence of the smaller values will be increased, and the value of the weighted average will be smaller than the value of the unweighted average.

2. If the larger weights are associated with the larger values and

²T. M. Adams, *Prices Paid by Farmers for Goods and Services and Received by Them for Farm Products, 1790–1871; Wages of Farm Labor, 1780–1937*, (A Preliminary Report from University of Vermont and State Agricultural College, Burlington, Vermont, February, 1939), pp. 43–44.

the smaller weights with the smaller values, the influence of the larger values will be increased and the resulting average will be larger than the unweighted average.

3. If there is no relation between the sizes of the weights and the values of the items, that is, if large weights are as frequently assigned to small items as to large items and small weights are similarly distributed, the difference between the weighted average and the unweighted average is likely to be very small and entirely due to chance; in fact under these circumstances the two averages may be the same.³

The computation of a weighted average can be explained from Table 73. The per capita sales in column 4 are the results of dividing each sales figure of column 3 by the corresponding population in column 1. The average per capita sales for the entire table is not obtained by averaging the figures in column 4 because the populations represented by the cities in the several size groups are different. The proper method is to multiply each per capita figure by the population of that group, sum the products, and divide by the total population. The result of this operation is \$97.79, as indicated. The same figure can be obtained by using as weights the percentage distribution of the population

TABLE 73
SALES OF RETAIL FOOD STORES (1935) AND POPULATION (1930)
IN CITIES OF DIFFERENT SIZE *

SIZE OF CITY	(1) POPULATION (2)		(3)	(4)
	Number (in thousands)	Percentage Distribution	RETAIL FOOD SALES (in millions)	PER CAPITA SALES
250,000 and over	28,785	31.2	\$2,839.3	\$ 98.64
75,000-250,000	9,771	10.6	941.5	96.36
10,000- 75,000	19,784	21.4	1,998.5	101.02
2,500- 10,000	10,615	11.5	1,175.2	110.71
Under 2,500 (excluding farms)	23,375	25.3	2,074.2	88.74
Total	92,330	100.	\$9,028.7	\$ 97.79

* Population, *United States Census*, 1930; Food store sales, *United States Census of Business*, 1935.

in column 2, and dividing the sum of the products by 100. The average per capita sales can also be computed by dividing the total sales by the total population, i.e., $\$9,028,700,000 \div 92,330,000 = \97.79 .

³ Cf. E. C. Rhodes, *Elementary Statistical Methods*. (London: George Routledge & Sons, Ltd., 1933), pp. 143-45.

When all of the necessary information is available as in this table, the weighted average should be computed as the ratio of the two pertinent totals. In most cases only the individual ratios will be given. Then weights must be found in order to compute the weighted average. The rule for determining these weights was stated in chapter XI in discussing the problem of averaging ratios.⁴ That development shows that in averaging the per capita figures the denominators from the computation of the individual per capita figures must be used as weights. But the average per capita sales in each size of city multiplied by the population in that group gives the total sales in the group and the sum of these products is the total sales. Therefore, the two methods just described for computing the weighted average are identical and one or the other should be used according to the form in which the data are available.

The use of the population column for weighting is required in order to retain in the weighted average a characteristic of the simple average, i.e., that the average times the number of items will give the total value of the original items. In the weighted average that rule becomes: the weighted average times the total of the absolute weights must equal the total value of the original items. In the example the average per capita sales times the total population equals the total sales. This characteristic will be referred to as the *total value criterion*.

The unweighted average of column 4 is \$99.10. The effect of the weights, therefore, has been to reduce the value of the average. This means that large weights tend to be associated with small per capita sales and small weights with large per capita sales. Survey of the table shows that, although the relation is somewhat mixed, the next to the largest weight is attached to the smallest per capita figure and the next to the smallest weight is attached to the largest per capita figure.

Another use of the weighted arithmetic average is found in the computation of the percentage changes in retail sales of independent dealers in Ohio. In this case the purpose is to derive from sample data an average figure that will represent the universe. Regular reports of current retail sales are collected monthly from a large number of co-operating merchants representing many lines of retail trade. By classifying and tabulating these reports according to the lines of trade which they represent, it is possible to calculate the percentage changes

⁴ See Chapter XI, pp. 263-69.

from the previous month in retail sales for all independent dealers in Ohio, as shown in Table 74.

TABLE 74

CALCULATION OF PERCENTAGE CHANGE IN RETAIL SALES BY INDEPENDENT DEALERS IN OHIO, USING WEIGHTED AVERAGE OF RELATIVES DERIVED FROM SAMPLE REPORTS REPRESENTING VARIOUS LINES OF TRADE, SEPTEMBER, 1939, DIVIDED BY AUGUST, 1939

LINE OF TRADE	(1) PERCENTAGE RELATIVES SEPT. 1939 ÷ AUG. 1939* (X)	(2) SALES OF EACH LINE AS RELATIVE OF TOTAL NET SALES 1935 CENSUS (W)	(3) WEIGHTED PERCENTAGE RELATIVES (1) X (2) (XW)
Grocery without meats.....	116.89	.0384	4.49
Grocery with meats.....	110.00	.1620	17.82
Country general	102.88	.0264	2.72
Department stores	125.45	.1367	17.15
Men's and boys' clothing.....	126.67	.0277	3.51
Family clothing	106.88	.0112	1.20
Women's ready-to-wear	123.54	.0260	3.21
Shoes	147.16	.0125	1.84
Motor vehicles	86.73	.2031	17.62
Gasoline filling stations.....	98.95	.0971	9.61
Furniture stores	100.51	.0380	3.82
Household appliances	122.43	.0123	1.51
Radio stores	108.30	.0026	.28
Lumber and building material....	95.71	.0297	2.84
Heating and plumbing.....	119.09	.0050	.60
Hardware	107.16	.0376	4.03
Restaurants	105.47	.0784	8.27
Drugs	100.32	.0381	3.82
Florists	111.29	.0055	.61
Jewelry	108.54	.0118	1.28
Total		1.000	106.23
Weighted percentage change....			+6.23
Unweighted percentage change†..	+15.18		

* Unpublished data.

† The unweighted percentage change is obtained by using the unweighted totals of reports from all lines of trade. The use of a weighting system with a total of 1.0 is explained in chapter XII.

It would be an easy matter to add the sales reported in all lines for the respective periods. The percentage change would then be the ratio between the two totals less 100 per cent. But this method, though simple, involves a basic error. The total sales of the reporting stores in each line of trade do not bear the same proportionate relationship to the total of all sales reported as the total value of these same lines of trade hold to the actual total value of all retail trade. This disproportionate relationship is due to the voluntary co-operative arrangement by which the collection process is carried on. In order to overcome this difficulty weights have been introduced. The percentage relatives,

column 1, calculated from the total of the reported sales of each line of trade, are multiplied by a number, column 2, representing the proportion which sales by independent dealers in each of the lines of retail trade constituted of the total of retail sales by independent dealers in Ohio as reported in the Census of Business of 1935.⁵ The sum of these products, column 3, becomes the basis for calculating the weighted percentage change. The effect of weighting is very important here, for the weighted change shows an increase of 6.23 per cent, whereas the unweighted change shows an increase of 15.18 per cent.

In this example the weights are not the denominators of the ratios to which the weights are applied. As a result, the weighted average times the sum of the absolute weights (the total sales upon which the relatives in column 2 are based) will not give the total value (in this case total sales reported in September, 1939). This departure from the established criterion is justified by the purpose of the computation, which is to estimate from the reported sample the percentage relation in a universe. The figure 106.23 is the best obtainable estimate of the percentage relation between the sales in September and August of all independent retail outlets in Ohio in the included lines of trade. That is, if it were possible to know the sales of all such stores in the state in August, the result of multiplying that figure by 1.0623 would approach the actual sales of these stores in September. This extension of the weighted average is frequently involved in estimating the conditions of a universe from those found in a sample.

Choice of Weights.—The problem of selecting weights does not arise so long as the total value criterion is adhered to. Two types of cases must be considered in which the criterion is abandoned. In both of these the question of choice of weights is a necessary preliminary step to the computation of the average. The principles involved in the choice are discussed in general terms here. A more specific discussion of the weighting problem in the construction of index numbers is presented in chapter XIX.

The first type of case arises when the conditions of a universe are to be inferred from a sample. The computation of the percentage change in retail trade in Ohio presented in the preceding section is an

⁵ It should be observed that the weights in this case represent the fractions which the annual volumes of the separate lines constituted of the total annual volume in 1935. To the extent that the different retail lines are differently affected by seasonal influences, these weights introduce errors. It is felt, however, that the error thus introduced is less than that due to lack of representative sampling.

example. The weights in this case represented the relative importance of the several lines of trade and were applied to percentage relations in the sample. This is a standard method of weighting to transfer from sample to universe, but by no means a universal one. Data in different form may require a different method of weighting. Therefore, no rule can be offered except the broad one that the weights must be so selected that the characteristics of the universe will be correctly inferred.

The second type of case appears when the true weights required to preserve the total value criterion are unavailable but some systematic weighting is an obvious necessity. The choice lies between approximating the missing true weights and adopting an arbitrary weighting system. The first alternative is employed in computing the average farm price of wheat at a particular time. A large number of individual prices are reported from various parts of the country but no corresponding reports of the amounts sold at various prices are available. Skillful estimators supply the missing quantity weights on the basis of whatever auxiliary knowledge can be gleaned from sources at their command. The result is a weighted average farm price closely approximating the correct average price. There is another reason for the success of this procedure in the hands of practiced estimators. Small variations in the weighting system will have comparatively little effect on the value of the average. For this reason Bowley stated as a general rule, "In calculating averages give all care to making the items free from bias, and do not strain after exactness of weighting."⁶

The second alternative, arbitrary weighting, is employed whenever an approximation to the true weighting system is not feasible. An example is found in the weights suggested on page 391 for determining an average monthly wage rate for farm labor. The weights attached to the several months are a composite of several elements and, since the total value criterion has been abandoned, the accuracy of the weighted result is largely dependent upon the judgment of the one who established the weighting system. The full meaning of judgment weighting is brought out in chapter XXV in connection with the construction of an index of business conditions from component series of unequal importance.

Weighted Total.—In some cases the weighted average is not as useful as the weighted total, i.e., $\Sigma (WX)$. For the most part this

⁶ Arthur L. Bowley, *Elements of Statistics* (London: P. S. King and Son, Ltd., 1920, fourth edition), p. 94.

applies when some standard unit of measurement exists independent of the particular investigation, and comparisons must be made in terms of that unit. For example, in calculating the cost of living of families for use in the administration of unemployment relief, the problem of family size arises at once. The number of persons per family is not a sufficiently accurate standard. Families of five persons, for instance, are not all equivalent, for the ages and sexes of the members of a family are very important in determining food requirements as well as in estimating clothing needs and housing costs. One approach to the solution of the problem of calculating food requirements which was presented by a group of experts under the auspices of the League of Nations involved assigning weights to persons according to their age and sex. A value of 1 was given to a male between 14 and 59 years of age and all other ages were assigned values relative to this. The scale of weights which was developed follows:

AGE AND SEX	WEIGHT
Under 2 years, male or female.....	.2
2- 3 years, male or female.....	.3
4- 5 years, male or female.....	.4
6- 7 years, male or female.....	.5
8- 9 years, male or female.....	.6
10-11 years, male or female.....	.7
12-13 years, male or female.....	.8
14-59 years, male.....	1.0
14-59 years, female.....	.8
60 years and over, male or female.....	.8

In order to obtain the weighted total of food units required for a family, it is necessary to assign the proper weight to the number of family members according to age and sex, and add the products. For instance, food units required for a family of five members, father aged 40, mother aged 35, one son aged 15, and two daughters, one aged 10 and one 11, is calculated as follows:

MEMBER	NUMBER	WEIGHT	TOTAL UNITS
Father, 40	1 male	1.0	1.0
Mother, 35	1 female	.8	.8
Son, 15	1 male	1.0	1.0
Daughters, 10 and 11.....	2 females	.7	1.4
Total	5 members		4.2

The weighted total of 4.2 represents the total number of food units required for the family in question. This is an average of .84 food

units per person for this family. Since the natural unit for relief is the family and not the individual, the weighted total is in this case a more useful figure than the weighted average. It should be observed that this system of weighting did not differentiate food requirements of the different sexes under 14 years of age or 60 years and over.

The Average of a Frequency Distribution

Frequency distributions are so commonly used in analyzing and describing various kinds of business data that it is necessary to examine the methods by which an arithmetic average can be calculated from such a distribution. The method is very similar to that employed in obtaining a weighted average. Each midpoint of a class interval is multiplied by the frequency of that interval. The sum of the products of midpoints multiplied by the frequencies represents the total value of all items in the distribution. When this total is divided by the sum of the frequencies, the result is the arithmetic average of the distribution.

Frequency distributions should be constructed so that the class marks represent all the values included in a class interval. Although this standard for construction may not always be attained, the class limits should be so established that in each class the midpoint of the class interval is approximately equal to the average of the actual values of the items. Each midpoint will therefore represent all the values in the class.⁷

TABLE 75
CALCULATION OF THE ARITHMETIC AVERAGE FROM THE FREQUENCY DISTRIBUTION
OF RENTALS PAID BY 155 FAMILIES IN COLUMBUS, OHIO

CLASS INTERVAL (dollars)	(1) CLASS MIDPOINT X	(2) FREQUENCY f	(3) FREQUENCY $\times$ MIDPOINT fX
7.50 and under 17.50	12.50	16	200.00
17.50 and under 27.50	22.50	27	607.50
27.50 and under 37.50	32.50	44	1,430.00
37.50 and under 47.50	42.50	17	722.50
47.50 and under 57.50	52.50	18	945.00
57.50 and under 67.50	62.50	11	687.50
67.50 and under 77.50	72.50	10	725.00
77.50 and under 87.50	82.50	9	742.50
87.50 and under 97.50	92.50	3	277.50
Total		155	6,337.50

$$M = \frac{6,337.50}{155} = \$40.89$$

⁷ See chapter XV for a complete discussion of the characteristics of frequency distributions which will affect the calculation of arithmetic averages.

Direct Calculation.—The arithmetic average of the rent distribution which was constructed according to this principle is calculated in Table 75 to illustrate the method. Since \$12.50, the midpoint of the first class interval, \$7.50 to \$17.50, is assumed to represent the average rental paid by the 16 families whose rentals fall within the interval, the total amount paid by the 16 families should be \$12.50 multiplied by 16, or \$200.00. Likewise, each product in column 3 represents the total rental paid by the families in that class interval of rentals. The total of column 3, \$6,337.50, should therefore be approximately equal to the total amount paid for rent by all families in the sample. The arithmetic average is found to be \$40.89 by dividing this total by 155, the number of families.⁸

It will be observed at once that \$40.89 differs slightly from the figure that was obtained by the computation of the arithmetic average of the original data on page 389, and the computed total rentals paid, \$6,337.50, is likewise a different total. A question arises at once as to which average or total ought to be used. Obviously the computations from the original data are more precise, but whether they should therefore be used will depend upon the purpose. The availability of the original data may also be a determining factor. On many occasions data are available only in frequency distributions so that there is no question as to which method of calculation to employ.

If the average and total were being used by a rental office in connection with its accounting records, the values of each separate rental would be on file and the simple arithmetic total and average would probably be computed from them. On the other hand, if the average of this sample is to be used to represent the average rent paid by all

⁸ It is frequently helpful to be able to describe the calculation of the average from the frequency distribution by a formula. Such a formula can easily be developed from the computation just completed. The procedure was as follows: Arithmetic average = (frequency of first class interval $\times$ midpoint of first class interval) + (frequency of second class interval $\times$ midpoint of second class interval) + etc., divided by the sum of the frequencies. Using the symbols at the head of each column in Table 75, the formula for the arithmetic average becomes:

$$M = \frac{f_1X_1 + f_2X_2 + \dots + f_nX_n}{f_1 + f_2 + \dots + f_n} \\ = \frac{\Sigma(fX)}{\Sigma f}$$

X is used to denote the values of the midpoints of class intervals in a frequency distribution just as X would be used to denote the individual values if the data were ungrouped

If a student understands the procedure in the calculation of the arithmetic mean, he need not memorize this formula. Note, however, that the denominator is always the total number of frequencies, *not* the number of classes in the distribution.

families in the city, calculation from the frequency distribution is preferable.⁹

The frequency distribution is a device for summarizing data and for reducing the amount of work involved in calculating statistical measures. When, as in the case of the rentals, a relatively small number of items is included in the distribution, differences may occur in the measures of central tendency calculated from the grouped and ungrouped data. As the size of the sample increases, however, differences in these calculated values tend to disappear.

Short-Cut Calculation.—The direct method of computing the arithmetic average from a frequency distribution is not an involved process, but the actual steps of multiplication, particularly in large distributions, may become a real burden. The number of computations can be reduced by the use of short-cut methods and as a consequence the chance for errors will be decreased. Although at a first glance, or upon an initial trial, these methods may not appear shorter, a little practice will convince anyone except an arithmetic wizard that much time and labor can be saved by employing them. They also lay the foundation for a much greater saving in more advanced calculations.

Method 1: The arithmetic average of rentals from the sample of 155 families is calculated by short-cut method 1 in Table 76.¹⁰

In carrying out the illustrative computations the first step is to select one of the midpoints¹¹ as an assumed average. Before calculation, of course, the average is not known, but for illustrative computation A the

⁹ See page 368 for discussion of the use of frequency distributions of sample data for representing the characteristics of a larger universe.

¹⁰ This text employs a fixed notation, the basis of which is as follows:

1. Capital letters ($X, Y \dots$) are used to denote values of variables measured from zero, e.g., the midpoints of class intervals in column 1, Table 76.

2. Small letters ($x, y \dots$) are used to denote values of variables measured from the average, e.g., the differences of the midpoints of the rent classes from the average rent, \$40.89.

3. The letter, d , is used to denote values of variables measured from an assumed average, e.g., columns 3 and 5, Table 76.

4. Primes and subscript letters will be used to distinguish variables that are being compared, e.g., d and d' in columns 3 and 5 to indicate deviations from two different assumed averages, and d_s to indicate deviations in steps in Table 77, column 3.

5. The measures of central tendency will be designated as follows:

M = true arithmetic average
 M' = assumed arithmetic average
 Me = median
 Mo = mode
 $G.M.$ = geometric average.

6. Additions to the notation will be made in subsequent chapters as the need arises.

¹¹ Any value can be used as the assumed average, but it has become customary to use the midpoint of a class interval because it affords easiest computation.

TABLE 76
SHORT-CUT METHOD 1 FOR COMPUTING THE ARITHMETIC AVERAGE
FROM THE FREQUENCY DISTRIBUTION OF RENTALS PAID BY
155 FAMILIES IN COLUMBUS, OHIO

CLASS INTERVAL (dollars)	(1) CLASS MID- POINT X	(2) FRE- QUENCY f	ILLUSTRATIVE COMPUTATION A		ILLUSTRATIVE COMPUTATION B	
			(3) Dollar Deviation of Midpoint from Assumed Average of \$42.50 d	(4) Frequency $\times$ Deviation (2) $\times$ (3) fd	(5) Dollar Deviation of Midpoint from Assumed Average of \$22.50 d'	(6) Frequency $\times$ Deviation (2) $\times$ (5) fd'
7.50 and under 17.50...	12.50	16	—30	—480	—10	—160
17.50 and under 27.50...	22.50	27	—20	—540	0	0
27.50 and under 37.50...	32.50	44	—10	—440	+10	+440
37.50 and under 47.50...	42.50	17	0	0	+20	+340
47.50 and under 57.50...	52.50	18	+10	+180	+30	+540
57.50 and under 67.50...	62.50	11	+20	+220	+40	+440
67.50 and under 77.50...	72.50	10	+30	+300	+50	+500
77.50 and under 87.50...	82.50	9	+40	+360	+60	+540
87.50 and under 97.50...	92.50	3	+50	+150	+70	+210
Total		155	...	—250	...	+2,850

Illustrative Computation A:

$$M' = 42.50$$

$$M = 42.50 + (-250 \div 155)$$

$$= 42.50 - (+250 \div 155)$$

$$= 42.50 - 1.61$$

$$= \$40.89$$

Illustrative Computation B:

$$M' = 22.50$$

$$M = 22.50 + (2,850 \div 155)$$

$$= 22.50 + 18.39$$

$$= \$40.89$$

midpoint \$42.50 is chosen as the assumed average. The interval for which the midpoint is \$32.50 is \$10 less than the assumed average, and so is shown in column 3 of Table 76 as deviating from the assumed average by — \$10; the midpoint \$22.50 is shown as \$20 less than \$42.50; the midpoint at \$52.50 is shown as \$10 more than the assumed average, etc. These differences are called actual dollar deviations of the midpoints from the assumed average, and are shown in order in column 3. The deviations are multiplied by their respective frequencies, and the products, retaining the algebraic signs of the deviations, are shown in column 4. These products are the amounts of difference between the total rentals actually paid in each class and the rentals that would have been paid if everyone had paid a rental equal to the assumed average. For instance, the 16 families in the first class interval actually paid \$480 less than they would have paid if each of the 16 families had paid \$42.50 per month in rent. The net total of column 4

indicates that the whole group of 155 families actually paid \$250 less in rent than they would have paid if everyone had paid \$42.50, the amount of the assumed average. With this information at hand, the arithmetic average can be computed as follows: Arithmetic average = assumed average + the net difference divided equally among all the items included (prorated net difference), i.e.,

$$M = 42.50 + (-250 \div 155) = 42.50 - 1.61 = \$40.89$$

This is the same average that was obtained by the direct method in Table 75. Illustrative computation B, at the right of Table 76, columns 5 and 6, shows that the assumed average can be taken at a different midpoint with identical results.

Following the same procedure as was employed in the direct method, a formula for short-cut method 1 can be developed:

$$M = M' + \frac{\Sigma(fd)}{\Sigma f}$$

Method 2: This method is a modification of method 1. The same deviations in actual amounts are used, but instead of being taken as actual values they are counted as equal "steps" of deviation from the assumed average. For instance, in this case each step is defined as equal to \$10. The calculation is then made as in Table 77.

The width of the class interval is conveniently chosen as the step in distributions having equal class intervals. The midpoint \$42.50 is again chosen as the assumed average. Each \$10 of deviation of a class interval midpoint from the assumed average is considered as one step, as in column 3. The multiplication of frequencies by these step deviations is done just as it was in the former illustration. For instance, the sum of the products of the frequencies and step deviations, column 4, indicates that the 155 families taken all together paid the equivalent of 25 steps (each \$10 wide) less in rent than they would have paid if each rental had been equal to the assumed average. The value of these 25 steps must now be prorated among the 155 rentals paid and reduced to dollar figures. The calculation is as shown below Table 77. The formula for this method can be written as:

$$M = M' + \frac{\Sigma(fd_i)}{\Sigma f}$$

in which i = the width of the step (usually the class interval) expressed in the original units. The chief advantage of this method over short-cut method 1 is that the multiplications are so reduced in size

that they can usually be performed mentally. In computing the average, the final multiplication of the prorated net difference by the width of the step must never be overlooked.

Frequency Distributions with Unequal Classes or Open Ends.—On some occasions, it is necessary to compute an arithmetic average from a frequency distribution in which the class intervals are unequal in width or which contains open intervals at either end. An open-end frequency distribution is one in which the lower limiting value or the

TABLE 77
SHORT-CUT METHOD 2 FOR COMPUTING THE ARITHMETIC AVERAGE BY STEP DEVIATIONS
FROM THE FREQUENCY DISTRIBUTION OF RENTALS PAID BY
155 FAMILIES IN COLUMBUS, OHIO

CLASS INTERVAL (dollars)	(1) CLASS MIDPOINT X	(2) FRE- QUENCY f	(3) DEVIATION IN STEPS FROM ASSUMED AVERAGE d_s	(4) FREQUENCY $\times$ DEVIATION IN STEPS (2) $\times$ (3) fd_s
7.50 and under 17.50...	12.50	16	— 3	— 48
17.50 and under 27.50...	22.50	27	— 2	— 54
27.50 and under 37.50...	32.50	44	— 1	— 44
37.50 and under 47.50...	42.50	17	0	0
47.50 and under 57.50...	52.50	18	+ 1	+ 18
57.50 and under 67.50...	62.50	11	+ 2	+ 22
67.50 and under 77.50...	72.50	10	+ 3	+ 30
77.50 and under 87.50...	82.50	9	+ 4	+ 36
87.50 and under 97.50...	92.50	3	+ 5	+ 15
Total		155	...	— 25

$$\begin{aligned}
 M' &= 42.50; i = 10.00 \\
 M &= 42.50 + (-25 \div 155) 10 \\
 &= 42.50 + (-.161) 10 \\
 &= 42.50 - 1.61 \\
 &= \$40.89
 \end{aligned}$$

upper limiting value, or both, is not indicated. Although open ends should be avoided whenever possible, if it is felt that such a class is necessary the total value of all the items in the class, or their average value, or the value of each individual item should be given in a footnote to aid in the description and analysis of the distribution. Unless information is supplied in one of these three forms, it becomes impossible satisfactorily to calculate the arithmetic average of the distribution, because the data required in the computation are not all available.

Table 78 illustrates an open-end frequency distribution that also contains unequal class intervals. The computation of the arithmetic average of this kind of distribution is shown in the table. The unequal class intervals in this case may be due to the administrative requirements of the association.

The arithmetic-average purchase can be calculated directly by dividing the total sales by the number of purchases, just as was done in the distribution with equal class intervals. Care must be exercised, however, to use the correct midpoints of the classes. Short-cut method 1 may save time in calculation, and, as shown in columns 4 and 5 of Table 78, can be applied in this type of distribution just as easily as in one with equal class intervals.

The step method is not commonly used in distributions with unequal classes. If it were employed in this calculation, columns 1, 2 and 4 would be unchanged. From column 4 it would be apparent that the width of the step should be \$5. The step deviations would read - 11, - 7, 0, + 11, + 32, and + 61. The computation would be,

$$M = \$65.00 - \frac{655}{724}(\$5) = \$60.48.$$

TABLE 78
METHOD OF CALCULATING THE AVERAGE VALUE OF PURCHASES OF ACTIVE PATRONS OF A
CO-OPERATIVE ASSOCIATION, JANUARY 1 TO DECEMBER 31, 1937*

VALUE OF PURCHASES (dollars)	(1) NUMBER OF PUR- CHASERS (FRE- QUENCY) <i>f</i>	(2) MID- POINT <i>X</i>	(3) FRE- QUENCY X MID- POINT <i>fX</i>	SHORT-CUT METHOD	
				(4) Deviation from Assumed Average, \$65.00 <i>d</i>	(5) Frequency X Deviation <i>fd</i>
0.00 and under 20.00..	248	10.00	2,480	- 55	-13,640
20.00 and under 40.00..	140	30.00	4,200	- 35	- 4,900
40.00 and under 90.00..	202	65.00	13,130	0	0
90.00 and under 150.00..	74	120.00	8,880	+ 55	+ 4,070
150.00 and under 300.00..	49	225.00	11,025	+160	+ 7,840
300.00 and over.....	11	370.00†	4,070	+305	+ 3,355
Total	724		43,785		- 3,275

$$M = \frac{43,785}{724}$$
$$= \$60.48$$

$$M = 65.00 + \left(\frac{-3,275}{724}\right)$$
$$= 65.00 - 4.52$$
$$= \$60.48$$

* From unpublished business reports of the association.
† Average value obtained by dividing total sales in the interval by the number of purchases.

The problem of the open-end distribution can be solved as indicated whenever the total of the open-end class is known, as it is in this case, or when the average value for the class is provided. Too frequently in published data neither of these values is known, so that, without employing dangerous assumptions, it becomes impossible to calculate the arithmetic average. Distributions of this type are common in many

kinds of census data. In such instances, it is necessary to employ one of the other measures of central tendency that depend upon position and do not make use of extreme values. These measures are described in chapter XVII.

THE GEOMETRIC AVERAGE

Definition

The geometric average is a measure of central tendency which, like the arithmetic average, depends upon the values of all the items in the group. It is formally defined as the positive value of the n th root of the product of n positive items. This definition may sound very forbidding, but the method of computation is relatively simple. In symbols, it becomes:

$$\text{Geometric mean} = \sqrt[n]{X_1 \times X_2 \times \dots \times X_n}$$

Following the definition it is only necessary to extract the n th root of the product of the n items included. This work is greatly facilitated by the use of logarithms, the geometric average being simply the anti-logarithm of the arithmetic mean of the logarithms of all the items.

The Average of Ungrouped Data

The use of logarithms in computing the geometric average is illustrated in Table 79. The logarithms come directly from the table in Appendix C. The sum of the logarithms divided by the number of items gives the logarithm of the geometric average, and the anti-logarithm is the geometric average. As shown at the bottom of the table the geometric average of these five numbers is 59.3, whereas the arithmetic average is 177.8. The arithmetic average is three times as great as the geometric average. This difference is the result of the greater importance of large values in the arithmetic average, which exceeds four of the five items. The geometric, on the other hand, has three items below it and two above. The example is intended to bring out the more representative character of the geometric as an average of values that are scattered as much as those in the table.

The question of representativeness is related to the properties of the two averages. The arithmetic average is so located that the sum of the deviations of the individual values from it will be zero. One value that greatly exceeds the others provides a large deviation that offsets

TABLE 79

COMPUTATION OF THE GEOMETRIC AVERAGE OF FIVE NUMBERS

NUMBERS	LOGS OF NUMBERS
6	.77815
22	1.34242
50	1.69897
175	2.24304
636	2.80346
<u>5)889</u>	<u>5)8.86604</u>
$M = 177.8$	$\log G.M. = 1.77321$
	$G.M. = 59.32$

many small ones near the point of concentration of the data. Thus whenever a few exceptionally large values are included in the set, the arithmetic average will exceed the values of a majority of the individual items.

The corresponding property of the geometric average is: the product of the ratios of the individual items to the average equals unity. In this computation an item one tenth as large as the average offsets an item ten times as large. For example, in Table 79 for the first item, 6, the ratio to the average is .101, and for the last item, 636, the ratio is 10.7. Therefore the two offset each other almost exactly in the computation of the average.

In general, the geometric average should be used when a few large items destroy the representativeness of the arithmetic average. This situation is particularly likely to arise in averaging ratios when most of them fall close to the lower limit of the available range, and a few have much higher values. The geometric average can frequently be employed to advantage in measuring average rates of change from one time period to another.

The Average of a Frequency Distribution

The geometric average can be computed from a frequency distribution by a method very similar to that employed in calculating the arithmetic average. It is necessary to remember the basic assumptions of the grouping of data in a frequency distribution: that all items in each class interval are evenly distributed throughout the interval, and that for each interval a single value must be selected which is representative of all the values in the interval or which is equivalent to the average of the values of the items in the interval. The midpoint of each inter-

val, which is equal to the arithmetic average of the class limits, was assumed to be the arithmetic mean of the values included in the interval. To be consistent in calculating the geometric average of a frequency distribution, the geometric average of the class limits should be used for this purpose, but the additional work involved in carrying out these calculations is not justified by the improvement in the results. As a consequence the class marks are assumed to be the geometric averages of the items in the several classes.

A formula for finding the geometric average of a frequency distribution can be developed from the corresponding arithmetic average formula by substituting logarithms of values for direct use of values, i.e.,

$$\log G.M. = \frac{\sum(f \log X)}{\sum f}$$

in which X stands for the class marks and f for the corresponding frequencies. The anti-logarithm of the results obtained by performing the operations indicated on the right side of the equation is the geometric average.

The steps in the process are illustrated in Table 80 by the computation of the average of the price relatives of 771 of the commodities included in the Bureau of Labor Statistics Index of Wholesale Prices. The relatives express the change produced in the United States price system by the outbreak of war in Europe. The average increase is 6.0 per cent. The significance of this change in a period of 30 days might be overlooked until one reflects that the prices in the table represent transactions approximating twenty billions of dollars monthly. Hence an increase of 6.0 per cent in price would add $1\frac{1}{4}$ billions of dollars to the exchange value of goods. The arithmetic average increase in the price relatives is 6.5 per cent. The difference between the two averages is far from negligible in terms of the increase in exchange value of goods.

In a distribution of ratios such as this, the arithmetic average places undue emphasis on the ratios above the peak of the distribution. This characteristic will be referred to in the discussion of index numbers as the inherent upward bias of the arithmetic average. A biased result can be avoided by the use of the geometric average. When distributions are approximately normal, however, this advantage of the geometric average tends to disappear because in such cases there will be very little difference between the two averages.

TABLE 80

CALCULATION OF GEOMETRIC AVERAGE FROM FREQUENCY DISTRIBUTION OF RELATIVES.
WHOLESALE PRICE CHANGES IN 771 COMMODITIES, FROM AUGUST
TO SEPTEMBER, 1939*

PRICE RELATIVES (per cents) SEPT. 1939 ÷ AUG. 1939	NO. OF COMMODITIES <i>f</i>	CLASS MARKS (per cents) <i>X</i>	LOG <i>X</i>	<i>f</i> LOG <i>X</i>
Less than 94.5	4†	...	...	7.66673§
94.5- 99.5	16	97	1.98677	31.78832
99.5-100.5	378	100	2.00000	756.00000
100.5-105.5	117	103	2.01284	235.50228
105.5-110.5	75	108	2.03342	152.50650
110.5-115.5	60	113	2.05308	123.18480
115.5-120.5	37	118	2.07188	76.65956
120.5-125.5	23	123	2.08991	48.06793
125.5-130.5	22	128	2.10721	46.35862
130.5-135.5	15	133	2.12385	31.85775
135.5-140.5	8	138	2.13988	17.11904
140.5-145.5	8	143	2.15534	17.24272
145.5-150.5	2	148	2.17026	4.34052
150.5-155.5	4	153	2.18469	8.73876
155.5 and over.....	2‡	...		4.41635§
Total	771			1561.44988

$$\log G.M. = \frac{1561.44988}{771}$$

$$= 2.02523$$

$$G.M. = 106.0 \text{ per cent}$$

* Prepared from Monthly Release, *Average Wholesale Prices and Index Numbers of Individual Commodities*, United States Bureau of Labor Statistics, August and September, 1939.

† 94, 92, 88, 61.

‡ 161, 162.

§ Sum of the logarithms of the individual relatives in this class.

Characteristics of the Geometric Average

Every value in a group of data must be included in the calculation of the geometric average and hence the value of this measure of central tendency cannot be influenced by individual judgment factors. In this respect, the geometric average and the arithmetic average are similar. They differ chiefly because in the geometric average small values have a greater effect than large ones, whereas the reverse is true in the arithmetic average. Consequently the arithmetic average is always larger than the geometric average.

The geometric average can be employed only when all the values are positive, and none is zero. In view of this restriction, a geometric average percentage of profits and losses, for instance, cannot be calculated, for such data include plus and minus values. And a geometric average of percentage decreases and increases can be calculated only by expressing them as percentage relatives.

The geometric average is not so easy to understand as the arithmetic average. Although the principle of the geometric average appears clear enough, its meaning is not easy to comprehend; computation and practice in interpretation are essential before it becomes a readily available tool in statistical analysis. The length of the computation and the lack of easily understood properties have been important factors in restricting its use as a measure of central tendency.

PROBLEMS

1. Explain the difference between an unweighted and a weighted arithmetic average.
2. Six averages are computed from a set of values for variables, X , and a set of weights, W , as follows:

A			B			C		
X	W	$X \times W$	X	W	$X \times W$	X	W	$X \times W$
7	10	70	24	10	240	5	1	5
24	1	24	12	2	24	15	10	150
5	12	60	7	4	28	24	12	288
15	2	30	5	12	60	7	2	14
12	4	48	15	1	15	12	4	48
5)63	29	232	5)63	29	367	5)63	29	505
12.6	8		12.6	12.7		12.6	17.4	

What is the relation of these averages to the discussion on pages 391–92?

3. One year ago an investor owned the following stocks and received the annual dividend returns as stated:

Stock	Investment	Dividend Return
A	\$ 5,000	\$300
B	12,000	480
C	2,000	160
Total	\$19,000	\$940

Average return 4.95 per cent

Today his stock holdings are as follows:

Stock	Investment	Dividend Return
A	\$ 8,000	\$480
B	6,000	240
C	5,000	400
Total	\$19,000	\$1,120

Average return 5.89 per cent

- a) How are the average rates of return obtained?
 - b) Inasmuch as none of the individual dividend rates has changed during the year, how do you explain the increase in the average return?
4. Compute the weighted average percentage of change in retail sales in the following lines of trade in Ohio in September, 1939, compared with August, 1939 (data from Table 74, page 394).

Grocery with meats	Gasoline filling stations
Department stores	Restaurants
Motor vehicles	Drugs

5. In accordance with the list on page 397 of the text compute the number of food units required for the following family:

	AGE
Husband	32
Wife	28
Grandmother	60
Son	5
Daughter	1

6. On a single graph, plot the following two distributions of earnings in Hosiery Mill XYZ:

WEEKLY EARNINGS (dollars)	SEMI-SKILLED EMPLOYEES	
	Male	Female
6.00 and under 10.00	0	21
10.00 and under 14.00	33	45
14.00 and under 18.00	91	56
18.00 and under 22.00	122	28
22.00 and under 26.00	74	12
26.00 and under 30.00	24	6
30.00 and under 34.00	4	0
34.00 and under 38.00	2	0
Total	350	168

7. a) Using data in Problem 6, find the average weekly earnings of (1) semi-skilled males, or of (2) semi-skilled females, using three different methods of computation. Indicate all computations.
 - b) From the shape of the distribution in Problem 6 and the value of the average in (1) or (2), whichever was assigned to you, state the characteristics of the distribution of earnings of either male or female hosiery workers.
8. a) (1) What is the arithmetic average income of the total families

- (3) Does the difference between the two averages indicate what the average annual cost of owning an automobile ought to be? Discuss.
- b) (1) Compute the percentage of families owning automobiles in each income group.
- (2) Compute the average of the percentages in (1).
- (3) 36,500 is what per cent of 68,200?
- (4) Does the result of either (2) or (3) give the percentage of families in this distribution that own automobiles? Explain.

CAR OWNERSHIP BY U. S. FAMILIES HAVING INCOMES LESS THAN \$5,000 BY INCOME GROUPS, DATA FROM A SURVEY IN 1933 IN 18 CITIES, BY THE U. S. DEPARTMENT OF COMMERCE

INCOME GROUP (dollars)	TOTAL NO. OF FAMILIES REPORTING	NO. OF FAMILIES OWNING CARS
0- 499	19,400	5,800
500- 999	15,800	7,200
1,000-1,499	13,700	8,600
1,500-1,999	9,300	6,700
2,000-2,999	7,000	5,600
3,000-4,999	3,000	2,600
Total	68,200	36,500

9. The number of stores operated by each of eight retail variety chains in 1936 was:

CHAINS	NO. OF STORES, Nov. 1936*
W. T. Grant & Co.....	477
H. L. Green Co., Inc.....	134
S. S. Kresge Co.....	731
S. H. Kress & Co.....	235
McCrary Stores Corp.....	194
G. C. Murphy Co.....	194
J. C. Penney Co.....	1,496
F. W. Woolworth Co.....	1,995

* *Survey of Current Business*, January, 1937.

- a) Compute the arithmetic average number of stores per chain.
- b) Compute the geometric average number of stores per chain.
- c) Explain why the geometric average is superior for these data.
10. a) Using the data in columns (A), (B), (C), (D), and (E) of Table 55, page 294, compute an unweighted geometric average index for each year.
- b) Compare your results with the weighted arithmetic average index appearing at the lower right of the table.
- c) Discuss the merits of each index as a measure of the credit standing of a prospective borrower.

11. a) From the following table, compute the arithmetic average and the geometric average:

PER CENT DISTRIBUTION OF INDUSTRIAL ESTABLISHMENTS IN THE UNITED STATES, ACCORDING TO VALUE OF PRODUCTS, 1925

VALUE OF PRODUCTS (dollars)		PERCENTAGE OF TOTAL ESTABLISHMENTS
5,000—	20,000	30
20,000—	100,000	37
100,000—	500,000	22
500,000—	1,000,000	5
1,000,000—	2,000,000	6
Total		100

- b) Which average is more representative for the data? Give reasons.

CHAPTER XVII

MEASURES OF CENTRAL TENDENCY—AVERAGES OF POSITION

THE preceding chapter was devoted to the discussion of those measures of central tendency that depend upon calculation processes. This chapter proceeds with the description of measures that are determined by their positions in a given set of data and hence require the exercise of the computer's judgment. There are two commonly used measures of this type, (1) the median and (2) the mode.

THE MEDIAN

Whether in an array or a frequency distribution, the median is the value of the middle item. Expressed more precisely, it is the value at the point on either side of which there is an equal number of items. That is, when the number of items is uneven the median has the value of the middle item; when the number is even it lies half way between the two items at the center.

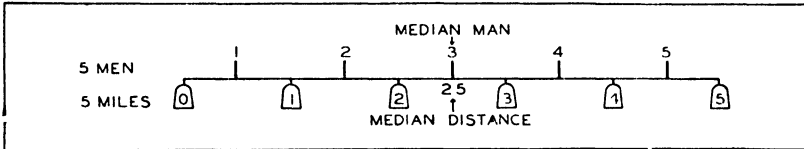
Location and Value of the Median in an Array

For a given set of data the location of the value of the median will be at the same item or between the same two items, whether the value of each item is known separately or whether they are grouped in a frequency distribution. Slightly different procedures are needed in the two cases, however, to fix the location of the median item and to determine its value. A simple diagram will explain the reason for this difference.

Suppose that we wish to find the middle point of a distance of 5 miles. Obviously it is located at 2.5 miles, computed by taking $\frac{5}{2}$, or $\frac{N}{2}$ if N = the number of miles. But if we have 5 men, one located at the center of each mile, the middle man is not the 2.5th but the 3d. From Figure 59, it is clear that each man's number is .5 more than the mileage at the point where he is standing. This is because the men are located and numbered at the *center* of each space, whereas each milestone is at the *end* of the space it measures. Thus

the 3d man stands at the 2.5th mile and the two medians coincide although they are designated differently.

FIGURE 59
LOCATION OF THE MEDIAN IN AN ARRAY



Application of Formula.—When individual items are arrayed they correspond to the 5 men in Figure 59. Therefore in order to find the middle item it is necessary to add .5 of an item to the fraction $\frac{N}{2}$. This is accomplished by adding 1 to the numerator. Thus the formula for locating the median item in an array is $\frac{N+1}{2}$. Its value is then simply that of the middle item or the average of the two central items.

The usefulness of the formula in finding the middle one of a large number of items is illustrated in Table 81. The formula for the median item is:

$$\frac{N+1}{2} = \frac{65+1}{2} = 33$$

Counting from either end of the array, the value of the 33d wage, \$53.35, is the median.

If there had been 64 pay checks in the group (omitting the last one, \$80.12) the solution by the formula would show the median to be the 32.5th item. It would then be a value half way between the 32d and 33d items and would be equal to one-half the sum (the arithmetic average) of the values of the 32d and 33d wages, i.e., $(\$52.93 + \$53.35) \div 2 = \$53.14 =$ the median. This value (\$53.14) is not the value of any single wage in the array, but is the value half way between the two central items and on each side of which there is an equal number of items.

The Extended Median.—It is obvious that a measure that is determined by the value of a single item or the average of two items has some unreliability, especially if there is only a small number of items in a sample. Suppose that an array consists of the five items, 5, 7, 15, 17, and 18; the median is 15. If two lower values, 2 and 3, are added, the median is 7; it has been reduced by 8 units, whereas if two higher values, 20 and 24, had been added instead, it would have been

TABLE 81

ARRAY OF WEEKLY WAGES OF 65 SKILLED RUBBER WORKERS
IN A FACTORY IN MICHIGAN*

\$36.96	\$31.92	\$53.35 (median)	\$56.43
38.84	51.93	53.43	58.34
38.96	51.96	53.58	58.58
41.12	52.48	53.66	59.13
47.02	52.62	53.73	60.01
47.99	52.73	53.93	60.12
49.07	52.73	54.32	61.36
49.29	52.73	54.74	62.69
49.35	52.73	55.52	63.45
49.43	52.73	56.31	68.62
49.43	52.83	56.43	68.62
49.43	52.83	56.43	71.34
50.03	52.83	56.43	73.42
51.16	52.83	56.43	73.49
51.90	52.92	56.43	73.49
51.92	52.93	56.43	78.82
			80.12

* Confidential unpublished source.

increased by only 2 units, to 17. In a small group therefore, unless the values are closely concentrated, the central item may be shifted so much by the addition of a very few items at either end that the median is too erratic to be depended upon. In such a case it is possible to resort to a more stable measure, the extended median. This is obtained by averaging the 3, 4, or 5 central items instead of taking a single item or the average of only two of them. Using the same example that was suggested above, the extended median of 5, 7, 15, 17, and 18 is $\frac{7+15+17}{3} = 13$. Adding the two lower values, the extended median of 2, 3, 5, 7, 15, 17, and 18 is $\frac{5+7+15}{3} = 9$; adding the two higher values, the extended median of 5, 7, 15, 17, 18, 20, and 24 is $\frac{15+17+18}{3} = 16\frac{2}{3}$. Thus the extended median fluctuated 4 points on one side of the center and $3\frac{2}{3}$ points on the other side when two more items were added at the ends of the series, whereas the ordinary median fluctuated 8 points on one side and 2 on the other. As the number of items averaged at the center is increased, the fluctuations will become smaller and more even on either side.

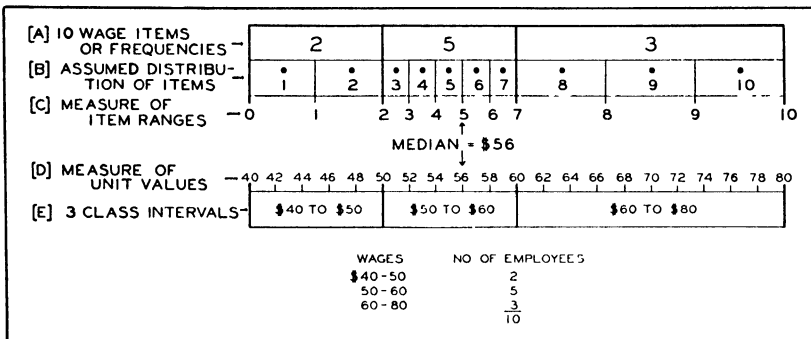
This measure is particularly valuable in determining the seasonal pattern of a time series, a purpose for which the median value of each month over a period of years is very commonly used. Speaking of the use of the extended median for obtaining seasonal variations, Christopher Saunders, of the University of Manchester, has said: "The extended median is probably the best average to take, since it is not influenced by extreme values which may be due to exceptional circum-

stances, and at the same time is not, like the simple median, affected by the accident of the value in any single year.”¹

Location and Value of the Median in a Frequency Distribution

In order to explain why the determination of the median in a frequency distribution differs from that in an array, it is necessary to go back to Figure 59. The 5 miles correspond to the range of unit values of class intervals in a frequency distribution. But instead of having one item in the center of each mile, or interval, there may be any number of items. In the absence of detailed information to the contrary, these are assumed to be evenly distributed throughout the interval, each one being at the center of its own “item range.” Figure 60 illustrates a very simple frequency distribution made up of

FIGURE 60
LOCATION OF THE MEDIAN IN A FREQUENCY DISTRIBUTION



10 weekly wages selected from Table 81 and grouped in three class intervals. If we knew the value of each of these 10 items, we could determine the location of the median from the formula for an array, $\frac{N+1}{2} = \frac{10+1}{2} = 5.5$, and its value would be half way between the 5th and 6th items (row B). But, since we know nothing except the class values and the number of items in each class, some other method must be found for determining an approximate value for the median.

Application of Formula.—The logical procedure is to interpolate from the assumed distribution of items in the class in which the median item falls to the unit values of that class. The median is still the value midway between the 5th and 6th items but, in order to interpolate on the diagram, the values of the items as well as of the

¹ Christopher Saunders, “Seasonal Variations in Employment,” *The Economic Journal*, Vol. XLV, No. 178 (June 1935), p. 272.

class intervals must be measured along a scale. Hence we use the "milestones" (row C) that mark the ends of the "item ranges"² instead of the numbers (row B) that count the items. The center of the 10 item ranges is therefore $\frac{N}{2} = \frac{10}{2} = 5$, and it will be seen from Figure 60 that this point in row C coincides with the point midway between the 5th and 6th items in row B. A line drawn from the end of the 5th item range intersects the scale of unit values (row D) at \$56, which is therefore the value of the median wage for this group of 10 items.

By computation, the class in which the median item will fall is determined by cumulating the frequencies until they exceed the value of $\frac{N}{2}$. In this case $\frac{N}{2} = 5$. This exceeds 2, the number of frequencies in the lowest class interval, but is less than $2 + 5$, the sum of the frequencies in the first two classes; hence the median falls in the 2d class, \$50 to \$60. The proportionate distance of the median value from the lower limit of the median class interval will be the same as the number of its item range in the interval is to the total number of item ranges in that interval. To determine which item range in the interval contains the median, subtract the sum of the frequencies below the median class from $\frac{N}{2}$. Thus, $5 - 2 = 3$, and the median value therefore is at the end of the 3d item range in the \$50 to \$60 interval. If x = the proportion of the \$10 interval, \$50 to \$60, that lies between \$50 and the median value, then

$$\begin{aligned}x : 10 &= 3 : 5 \\5x &= 30 \\x &= 6\end{aligned}$$

The median = \$50 + \$6 = \$56, which is the same result that was reached by use of the diagram. In general terms:

Me = the value of the median

m_1 = the number of item ranges Me is above the lower limit of the median class

m_2 = the number of item ranges Me is below the upper limit of the median class

f = the number of items in the median class

L = the lower limit of the median class

U = the upper limit of the median class

$$Me = \frac{m_1 U + m_2 L}{m_1 + m_2}$$

² The term "item range" refers to the portion of each class interval occupied by a single item, on the assumption that the items are evenly distributed within each class. In this example, since there are five items in the median class, each is assumed to occupy $\frac{1}{5}$ of the interval.

This formula can be used equally well in approaching the median from either end of the table by remembering that $m_1 + m_2 = f$.³

Thus in the example

$$m_1 = 3 \text{ and } m_2 = 2$$

and

$$Me = \frac{(3 \times 60) + (2 \times 50)}{3 + 2} = \frac{280}{5} = \$56$$

In order to avoid a confused diagram the example used in developing the meaning of item ranges was made very simple. The advantage of the resulting formula will be more evident when it is applied to a larger frequency distribution such as the rent data of Table 82. The

TABLE 82
FREQUENCY DISTRIBUTION OF RENTALS PAID BY 155 FAMILIES IN COLUMBUS, OHIO
ARRANGEMENT FOR CALCULATING THE MEDIAN

CLASS INTERVAL (dollars) (i = 10.00)		(1) FREQUENCY <i>f</i>	(2) CUMULATED UPWARD	(3) CUMULATED DOWNWARD
7.50 and under	17.50.....	16	16	..
17.50 and under	27.50.....	27	43	..
27.50 and under	37.50.....	44 median class	..	..
37.50 and under	47.50.....	17	..	68
47.50 and under	57.50.....	18	..	51
57.50 and under	67.50.....	11	..	33
67.50 and under	77.50.....	10	..	22
77.50 and under	87.50.....	9	..	12
87.50 and under	97.50.....	3	..	3
Total		155	...	...

median item range is $\frac{155}{2} = 77.5$. There are 43 items in classes preceding the median class, column 2, therefore $m_1 = 77.5 - 43$ or 34.5, and $m_2 = 44 - 34.5$ or 9.5. The value of m_2 can be verified from the upper end of the distribution, $77.5 - 68$, column 3, = 9.5. Then

$$Me = \frac{(34.5 \times 37.50) + (9.5 \times 27.50)}{44} = \frac{1,293.75 + 261.25}{44} = \frac{1,555}{44} = \$35.34$$

The definition of the median implies that the middle item can be located and its value calculated with ease irrespective of the arrangement of the frequencies in the distribution. Open or closed ends are

³ This is the usual form for straight line interpolation. It is equivalent to the formula sometimes used in computing the median:

$$Me = L + \frac{m_1}{f} i \quad \text{or} \quad Me = U - \frac{m_2}{f} i$$

where $i = (U - L)$, the width of the median class interval.

likewise immaterial so far as the calculation of the median is concerned. In fact, of all the measures of central tendency the median is least affected by the peculiarities of frequency distributions. The formula just developed for computing it may be employed with any frequency distribution. It is applied for illustration to the distribution in Table 83, which has unequal class intervals and open ends. The frequencies shown in column 1 are used in the computation.

TABLE 83
NUMBER OF INDIVIDUAL INCOME TAX RETURNS BY NET INCOME CLASS, 1930*

NET INCOME BEFORE PERSONAL DEDUCTION (dollars)	(1) NUMBER OF RETURNS (in thousands)	(2) NUMBER OF RETURNS (rounded to nearest fifty thousand)
Under 1,000.....	150	150
1,000 and under 2,000.....	910	900
2,000 and under 3,000.....	780	800
3,000 and under 5,000.....	1,070	1,050
5,000 and under 10,000.....	550	550
10,000 and over.....	260	250
Total	3,720	3,700

* *Statistical Abstract* (1939), p. 186.

$$\frac{N}{2} = 1,860; m_1 = 20; m_2 = 1,050; L = 3,000; U = 5,000$$

and

$$Me = \frac{(20 \times 5,000) + (1,050 \times 3,000)}{1,070} = \$3,037$$

It should be noted that the unequal class intervals and the open-end intervals caused no difficulty whatever in the use of the formula. The only case that might give rise to some confusion is when the median falls exactly at a class limit. This situation can be illustrated by rounding off the frequencies of column 1 of Table 83 to the nearest 50,000 returns as shown in column 2. The median item range is now the 1,850th one. But $150 + 900 + 800 = 1,850$, hence the median falls at \$3,000. The formula is sufficiently general to include this case: if the third class interval is taken as the median class, m_2 is zero; if the fourth class interval is taken, m_1 is zero. Either substitution will give \$3,000 as the median.

Graphic Calculation.—The value of the median can be determined with a fair degree of accuracy from the ogive. The process is simple and can be explained quickly with the aid of a diagram. The graph of the cumulative distribution of rents presented on page 373, Fig-

ure 54, contains a "less than" ogive and a "more than" ogive. The intersection of the two curves is at the 50 per cent ordinate on the right vertical scale; hence a perpendicular from this intersection meets the base line at the approximate value of the median. From the graph the median value appears to be about \$35.25, which coincides quite closely with the calculated value, \$35.34.

The value of the median can be obtained equally well from a diagram containing only one ogive. A horizontal line is drawn from the 50 per cent ordinate of the vertical scale until it intersects the ogive. The perpendicular dropped from this intersection will meet the base line at the median value. If the percentage scale were omitted from the diagram, the horizontal line would be drawn from the ordinate 77.5 on the left scale until it intersected the ogive. The intersection would, of course, be identical with the previous one.

Graphic calculation of the median is less precise than the use of the formula. However, the value of the median calculated from a frequency distribution is always an interpolated value; hence at times the greater precision of the arithmetic process may be spurious. In such cases the graphic method will give as much precision as warranted by the data.

THE MODE

The mode in statistics means exactly what it does in the dictionary: the prevalent or most frequently encountered thing. The concept of the mode is a part of everyday speech, but usually without the arithmetic precision that would cause one to think of it as a statistical measure. Illustrative of this use is the response by a clothing salesman to the question, "What color are men wearing this season?" "We sell more blue than any other color." Or to the question asking what amount a person should contribute to a charity, solicitors are frequently heard to say, "Most of my contributions have been between \$5 and \$10." In either of these examples a more exact statement of the mode could be made from a detailed tabulation of the data, but the purpose is merely to express a general impression based on the speaker's experience.

The modal wage is the one received by the greatest number of workers. The modal rent is that one paid most frequently. The modal weight of college students is the one that occurs more often than any other. These examples suggest immediately that a set

data must be obtained before the mode can be determined. Further examination indicates that the data must be carefully chosen. Wage data that included skilled and unskilled workers, men and women, or industrial and agricultural workers would be so heterogeneous that perhaps no modal wage could be found. Even though a modal wage appeared it could not be explained because of the diverse wage tendencies present in the data. Similar arguments apply to rent or weight data. We conclude therefore that the mode can be determined only from homogeneous data. This requirement is, of course, not peculiar to the mode, but is restated here because the difficulties that arise in determining the mode are in many instances directly traceable to the use of non-homogeneous data.

Attention must also be given to the form in which a set of data should be arranged for determining the mode. The array immediately comes to mind as a convenient arrangement. For example, in the array of weekly wages of Table 81 the most frequently occurring wage is \$56.43. There are other amounts, however, such as \$52.73 and \$52.83, that appear repeatedly and cause some doubt as to whether the most frequently occurring wage is the typical one. Grouping the wages in five-dollar intervals shows that the greatest concentration occurs between \$50 and \$55. In the light of this example, which is typical of what may occur in any array, it is probably best to set up a general operating rule that the mode should be determined only from a grouped frequency distribution.³

Even in a frequency distribution an exact value of the mode cannot be determined by elementary methods. Since the exact development lies outside the scope of this text, a method of finding an approximate value of the mode will be explained here. The technique differs slightly according to whether the class intervals of the distribution are equal or unequal. The two cases will accordingly be considered separately.

Frequency Distribution with Equal Class Intervals

It is possible by inspection to select the class interval in which the greatest number of items occurs. By interpolation within this class interval a value can be determined which is considered to be the mode.

³ A rather obvious exception to this operating rule occurs when in an array such a large part of the cases have a single value that no process of grouping could dislodge this value from its modal position

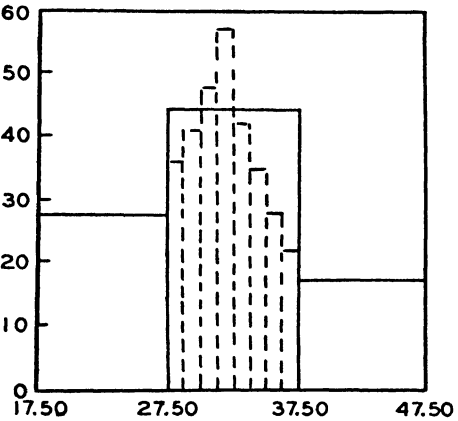
The basis for the method of interpolation and the computations involved in carrying it out will be illustrated by the use of the sample of rents in Columbus, as shown in Table 84. The class interval containing the greatest frequency (called the modal class) is the third

TABLE 84
FREQUENCY DISTRIBUTION OF RENTALS PAID BY 155 FAMILIES IN COLUMBUS, OHIO
ARRANGEMENT FOR CALCULATING THE MODE

CLASS INTERVAL (dollars)	(1) FRE- QUENCY <i>f</i>	(2) FREQUENCIES IN ONE CLASS ABOVE AND BELOW MODAL CLASS	(3) FREQUENCIES IN TWO CLASSES ABOVE AND BELOW MODAL CLASS
7.50 and less than 17.50.....	16	..	16
17.50 and less than 27.50.....	27	27 below	27 (43) below
27.50 and less than 37.50.....	44 modal class	..	..
37.50 and less than 47.50.....	17	17 above	17 (35) above
47.50 and less than 57.50.....	18	..	18 (35) above
57.50 and less than 67.50	11	..	..
67.50 and less than 77.50.....	10	..	..
77.50 and less than 87.50.....	9	..	..
87.50 and less than 97.50.....	3	..	..
Total	155	44	78

one, with limits \$27.50 and \$37.50, in which there are 44 items. Let it be assumed that this class can be subdivided into many narrow classes and that a diagram is drawn to represent these smaller classes by columns whose heights are adjusted to give the same total area as the single column for the class in a histogram of the original distribution. If this were a distribution containing no artificial points of concentration, the tops of these narrow columns on the base \$27.50

to \$37.50 would naturally follow the characteristic shape of a frequency distribution, i.e., rise gradually to a peak somewhere within the interval and then recede. The dotted columns of the diagram at the right are a rough illustration of the preceding description. The actual heights of such columns could be obtained only from an original array. The illustration does not actually represent the original rent data, which contained artificial grouping at two points within the interval instead of at the center. In the diagram, therefore,



the contour of frequencies within the modal class is what might be expected in any continuous distribution containing no irregularities.

Usually the original array from which a given distribution has been constructed is not available; consequently the approximate contour of the intramodal class frequencies is inferred by a reasoning process. If the frequencies of the two classes adjacent to the modal class are equal it is assumed that the modal class frequencies will be distributed normally and the peak will fall at the center of the interval. The value of the mode will then be the class mark of the modal class. If the frequency of the class below the modal class (in value) exceeds the frequency of the class above (in value), it is reasonable to suppose that the frequency contour within the modal class will resemble that of the illustrative diagram, and consequently that the mode will fall below the center of the modal interval. A similar argument leads to the conclusion that the mode will exceed the class mark of the modal class when the frequency of the interval above is greater than the frequency of the interval below.

The frequency in the interval higher in value than the modal class tends to exert an upward influence on the value of the mode from the lower limit of the modal interval and the frequency in the class interval below the modal class tends to exert a downward influence from the upper limit of the modal interval. The total influence in both directions is measured by the sum of the frequencies in the two classes.

This description can be converted into the interpolation formula,⁴

$$Mo = \frac{(f_l \times L) + (f_u \times U)}{f_l + f_u}$$

in which

Mo = the value of the interpolated mode

L = the lower limit of the modal class interval

U = the upper limit of the modal class interval

f_l = the frequency in the class interval just less than the modal class

f_u = the frequency in the class interval just greater than the modal class

⁴ This formula is the equivalent of two that are sometimes used in finding the interpolated value of the mode. The formulas,

$$Mo = L + \frac{f_u}{f_l + f_u} z$$

and

$$Mo = U - \frac{f_l}{f_l + f_u} z,$$

in which

z = width of the modal class ($U - L$),

can be converted to the form in the text by substituting $(U - L)$ for z and simplifying.

From Table 84,

$$Mo = \frac{(27 \times 27.50) + (17 \times 37.50)}{27 + 17} = \frac{742.50 + 637.50}{44} = \frac{1,380}{44} = \$31.36$$

This formula can be expressed as follows: *The interpolated value of the mode is obtained by taking a weighted average of the upper and lower limits of the modal class, the frequencies of the corresponding adjacent classes being used as weights.*

If the frequencies in the class intervals just above and below the modal class reveal great asymmetry or indicate some peculiar condition of grouping that cannot be improved, the interpolated value of the mode should be calculated by using the sum of two equal-sized intervals on either side of the modal class. For instance, in the frequency distribution of rentals the frequencies in the two intervals below the modal class decrease very sharply from 27 to 16, while the frequencies in the two intervals above remain almost constant; in fact, the second interval above increases slightly from 17 to 18. In this case the formula would be changed only by the inclusion of the two additional classes and the form would be,

$$Mo = \frac{(f_{l_1} + f_{l_2})L + (f_{u_1} + f_{u_2})U}{(f_{l_1} + f_{l_2}) + (f_{u_1} + f_{u_2})}$$

the subscripts being used to distinguish the pairs of classes, i.e.

$$f_{l_1} = 27, \quad f_{l_2} = 16, \quad f_{u_1} = 17, \quad f_{u_2} = 18$$

and

$$Mo = \frac{(27 + 16) 27.50 + (17 + 18) 37.50}{27 + 16 + 17 + 18} = \frac{1,182.5 + 1,312.5}{78} = \frac{2,495}{78} = \$31.99$$

This value, it will be observed, is \$.63 larger than the interpolated mode calculated by the former method. This increase is due entirely to the peculiarity of the distribution of this sample of rentals, which results in a relatively larger frequency in the second class above the modal class interval than in the second class below it.

With justification any student at this point can ask, "Which of these values of the mode is right?" The answer must be: "There is never an exact value of the mode that is right, while all others are wrong." The choice of which value to use depends upon the judgment of the statistician after an analysis of all the characteristics of the data. A

working rule might be stated as follows: use the first method unless a study of the distribution reveals that peculiarities which will reduce the validity of the first method can be overcome by the use of the second, or some modification of it.

Another departure from the usual formula is employed in dealing with a distribution of the kind shown in Table 85. The modal class interval appears to be \$30 and under \$35. But closer inspection reveals that the frequencies occur in fairly symmetrical pairs on the two sides of the nearly equal center pair, 373 and 377. One could reasonably assume, therefore, that the mode ought to fall in the vicinity of \$30. The original formula for computing the mode can be adapted to this assumption by taking the modal class as \$25 to \$35 and proceeding as usual. The computation would be,

$$Mo = \frac{(25 \times 351) + (35 \times 306)}{351 + 306} = \frac{19,485}{657} = \$29.66$$

This method of determining the mode should be used only when a study of the data reveals two adjacent classes at the peak with approximately equal frequencies and corresponding classes on the two sides of the enlarged center interval.

TABLE 85
BIWEEKLY WAGES OF "HAND LOADERS" IN BITUMINOUS COAL MINES
IN ALABAMA, 1924 *

WAGES FOR TWO WEEKS (dollars)	NO. OF MINERS	WAGES FOR TWO WEEKS (dollars)	NO. OF MINERS
Under 5.....	135	50 and under 55.....	143
5 and under 10.....	186	55 and under 60.....	90
10 and under 15.....	192	60 and under 65.....	63
15 and under 20.....	263	65 and under 70.....	50
20 and under 25.....	351	70 and under 75.....	26
25 and under 30.....	373	75 and under 80.....	28
30 and under 35.....	377	80 and under 90.....	19
35 and under 40.....	306	90 and under 100.....	8
40 and under 45.....	228	100 and under 110.....	5
45 and under 50.....	213	110 and under 120.....	3

* *Bulletin No. 416*, United States Bureau of Labor Statistics, 1926.

Frequency Distribution with Unequal Class Intervals

Open end intervals, unless adjacent to the modal class, make no difference in the computation of the value of the mode, but unequal class intervals do affect its calculation. The methods just described for determining the mode make assumptions which must be faithfully observed. The widths of the class intervals above and below the modal

class must be equal in order that frequencies can exert their effects through an equal range of the value of the variable. If the intervals are not equal, an adjustment must be made in some way to equalize them. This adjustment may be made by reconstructing the distribution from the array. When reconstruction is not feasible, the alternative is to combine class intervals at the center of the distribution so that the modal class and the two adjacent to it will be equal in width. In some cases it may be necessary to regroup with the two adjacent classes wider than the modal class, but the reverse with the modal class wider than the adjacent ones should be avoided, except for the situation explained in Table 85.

The method of regrouping will be illustrated by the use of Table 83, column 2. As the class intervals are given, the modal class appears to run from \$3,000 to \$5,000 with a frequency of 1050. But this class has a width of \$2,000, whereas the two lower-valued classes \$1,000 to \$2,000 and \$2,000 to \$3,000 with nearly as great frequencies are only half as wide. If these two classes are combined the resulting class \$1,000 to \$3,000 is modal with 1,700 frequencies. But the class below \$1,000 is not the same width as the upper adjacent class, \$3,000 to \$5,000. Therefore any attempt to obtain a specific mode by interpolation can be only approximate. A method that follows judgment more than rules would be to use half of the frequency in the \$3,000 to \$5,000 class as an expression of the upward influence acting against the downward influence of the 150 items in the "Under \$1,000" class. The computation would be

$$Mo = \frac{(150 \times 1,000) + (525 \times 3,000)}{675} = \$2,556$$

This value could differ considerably from the true mode (a figure which can be determined only by complete analysis of a detailed distribution of tax returns) but it falls just above the minimum taxable amount for persons with dependents, i.e., at the point where taxable returns of persons with dependents as well as those without dependents first appear in the table. There is some reason, therefore, for believing that this value of the mode may fall very close to the point of the maximum density of tax returns.

Bi-Modal Distributions

Careful study of the characteristics of data and the closest attention to the problems of grouping do not always lead to a single mode in a

distribution. In some cases the tabulation or graph of a frequency distribution reveals the presence of two modes or two points of about equally high frequency. This kind of situation does not affect the calculation of the arithmetic mean or the median, but it causes difficulty in the determination of the value of the mode; in fact, in such a case a value for the mode cannot be obtained without some rearrangement of the data.

A bi-modal distribution may be the result of faulty grouping of the data, too few cases in a sample, or lack of homogeneity in the data. The method of eliminating bi-modality varies according to which of these causes appears to be responsible. If the grouping is faulty, the error can be corrected only by returning to the ungrouped data. If these are not available, the distribution might as well be abandoned, unless by chance the fault can be corrected by combining the classes in larger groups. Bi-modality due to insufficient data can sometimes be overcome by increasing the width of the class intervals. If lack of homogeneity of the data has caused the two points of concentration, the difficulty can be removed only by separating the given distribution into its several homogeneous parts.

The latter situation is illustrated in Table 86 by the distribution of sales in a grocery store. A peak number of sales is apparent at \$.70 to \$.80 and another peak at \$1.00 to \$1.10. Before computing the mode, the statistician inquired more closely into the sales policy of the store and discovered that delivery service was being provided free on all orders for \$1.00 or more but that a charge of ten cents per order was being made on all orders less than \$1.00. The total distribution, therefore, was really two distributions combined: one consisting of across-the-counter sales and the other of delivered sales. When these two parts were separated each was homogeneous within itself. The separate tabulations are shown in Table 87. The respective modes are \$.75 for over-the-counter sales and \$1.09 for delivered sales.⁵

This particular analysis became useful in determining whether the delivery service should be maintained or, if the service was continued, whether the 10-cent charge should be eliminated. Studies such as this, though infrequently made, can be related to the various practices and policies of small establishments and may be helpful in the administration of the enterprise.

⁵ Another example of a heterogeneous bi-modal distribution with an explanation of its analysis will be found in chapter XXX.

TABLE 86
DISTRIBUTION OF SALES OF A SERVICE GROCERY STORE ON A
CERTAIN TUESDAY IN 1939*

SIZE OF SALE (dollars)		NUMBER OF SALES
Under	.30.....	35
.30 and under	.40.....	28
.40 and under	.50.....	35
.50 and under	.60.....	42
.60 and under	.70.....	52
.70 and under	.80.....	72
.80 and under	.90.....	51
.90 and under	1.00.....	48
1.00 and under	1.10.....	75
1.10 and under	1.20.....	46
1.20 and under	1.30.....	34
1.30 and under	1.40.....	28
1.40 and under	1.50.....	19
1.50 and under	1.75.....	15
1.75 and under	2.00.....	12
2.00 and under	2.50.....	5
2.50 and over.....		6
Total		603

* Confidential unpublished source.

It should be noted that even though a single modal class could have been secured in the combined distribution by regrouping in wider intervals, the resulting mode would not have been representative of either type of sales. The only method of dealing with a non-homogeneous

TABLE 87
DISTRIBUTION OF GROCERY SALES OF TABLE 86, SEPARATED
INTO ACROSS-THE-COUNTER AND DELIVERED

SIZE OF SALE (dollars)		NO. OF ACROSS-THE-COUNTER SALES	NUMBER OF DELIVERED SALES
Under	.30.....	35	...
.30 and under	.40.....	28	...
.40 and under	.50.....	35	...
.50 and under	.60.....	42	...
.60 and under	.70.....	52	...
.70 and under	.80.....	72	...
.80 and under	.90.....	51	...
.90 and under	1.00.....	45	3
1.00 and under	1.10.....	35	40
1.10 and under	1.20.....	24	22
1.20 and under	1.30.....	15	19
1.30 and under	1.40.....	11	17
1.40 and under	1.50.....	7	12
1.50 and under	1.75.....	4	11
1.75 and under	2.00.....	2	10
2.00 and under	2.50.....	1	4
2.50 and over.....		2	4
Total		461	142

distribution is to separate it into homogeneous parts. If the original ungrouped data or separated subdistributions are not available, the statistician has no choice but to abandon the analysis of data that he knows to be heterogeneous.

CRITERIA FOR SELECTING AND JUDGING AVERAGES

Throughout this chapter and the preceding one the major emphasis has been placed on methods of computing four measures of central tendency. In the course of the several explanations the distinctive features of the measures have been set forth in more or less detail, but in incidental fashion. At this point the student has a right to ask, "Which of these various averages should I use?" or "When ought I to use one or the other of the averages described?" There is no arbitrary single answer that can be given to these questions. The selection of the proper measure of central tendency depends upon (1) the purposes for which the measure is to be used, (2) the nature of the data available, and (3) the characteristics of the different measures. Each statistician must answer (1) for himself. The preceding pages provide the basis for answering (2), for in them are described the calculation requirements and limitations of each of the measures discussed. The following summary of the characteristics of the various measures will aid in answering (3). Each measure of central tendency meets some of these criteria, but the distinctive features possessed by each measure are likely to be most important in determining which one to use in a particular case.

Rigidly Defined

The arithmetic average and the geometric average are computed from invariant formulas. The results obtained are completely independent of the person doing the work. The two are therefore said to be rigidly defined. The interpolated value of the median in a frequency distribution is rigidly defined, but judgment enters when an extended median is computed for ungrouped data. The exact value of the mode in perhaps a majority of practical applications depends upon the computer's judgment. He must decide which frequencies to use in the formula and in the case of unequal intervals even the determination of the modal class is a matter of judgment. The mode is therefore less rigidly defined than any of the other averages.

All Items Used

The arithmetic average and the geometric average make use of all of the items according to their value. The median also takes account of all items but only according to their position. The items are important by count only; an item above the center and one below have equal weight in the location of the median. It is not necessary that the classes be defined precisely except near the middle of the distribution. The mode places all of the emphasis on the shape of the distribution near the maximum frequency. Thus it is affected by the values of items near the point of greatest concentration and by the degree of concentration. It is not affected at all by either the value, the number, or the position of items beyond this concentrated area.

Affected by Extreme Items

The arithmetic average includes each item according to its value; hence high valued items exert the greatest influence on the average. A few very large items may lead to an unrepresentative value of the average. This feature will be important in the development of index numbers. Extremely small items are given more emphasis by the geometric average than by the arithmetic. Neither the median nor the mode gives any weight to the values of extreme items. They are, therefore, particularly well adapted for use in open-end distributions in which the extreme values are not exactly defined. In such cases exact determination of the averages of calculation is impossible.

Easily Comprehended

Measures that possess some obvious property are easy to comprehend. The "mode" is widely understood even by those not trained in statistics. The concept of the median, the value of the middle item of a set, is almost as easy to grasp. The arithmetic average is the center of gravity of a distribution. It is the value such that the algebraic sum of the deviations of individual values from it is zero. In this sense it is easy to comprehend. The geometric average possesses no such simple property and that fact explains in large measure its infrequent use in practical work even where the type of data and the purpose of an investigation indicate that it should be used.

Easily Calculated

This criterion includes questions of complexity of method, length of calculation, availability of short-cut methods of calculation and necessity for special arrangements of data. The method of computing the median is simple, the calculation is brief, the data seldom require rearrangement, except to put individual items in an array, and it could be said that no short cuts are needed. The arithmetic average involves a simple calculation but one that is sometimes fairly long in spite of available short cuts; rearrangement is unnecessary. The calculation of the mode is short although somewhat complex because of variable method. The greatest difficulty arises from the frequent necessity of rearranging the data in order to establish a modal class and two adjacent classes of equal width. The method of computing the geometric average is by comparison both long and complex due to the use of logarithms. No simple short-cuts are available.

Subject to Algebraic Treatment

Many kinds of statistical analysis, whether advanced or elementary, make use of algebraic processes. Therefore, if data are to be used in any subsequent analysis of this sort, it is necessary that the averages used should be subject to algebraic treatment.

The arithmetic average and geometric average meet this requirement, largely because of other characteristics which they possess, such as rigid definition and inclusion of all items. The median and the mode do not lend themselves to algebraic treatment; consequently they are seldom used in advanced analysis.

Not Affected by Chance Arrangement

The particular arrangement of the data may affect the calculation of the various measures of central tendency very differently. Ordinarily, it may be assumed, the particular way in which data arrange themselves or happen to fall should not affect the value of the measure. Arrangement of the data near the modal class may depend on the chance distribution of the frequencies, therefore the mode is most vulnerable on this score. If the median is to be determined from a small number of items, chance shifting near the center may move the position of the median considerably. The arithmetic average and the geometric average do not depend upon the position of the items; hence their values are less affected by shifting of items due to chance

Summary

In order to show which of the measures of central tendency have the characteristics just described, Figure 61 was developed. By its use a comparison of the characteristics of the various measures can be surveyed quickly.

FIGURE 61
SUMMARY OF CHARACTERISTICS OF MEASURES OF CENTRAL TENDENCY

CHARACTERISTICS	ARITH- METIC MEAN	MEDIAN	MODE	GEO- METRIC MEAN
1. Rigid definition	Yes	No	No	Yes
2. All items are necessary.....	Yes	No	No	Yes
3. Values of extreme items have no effect....	No	Yes	Yes	No
Values of large items receive emphasis...	Yes	No	No	No
Values of small items receive emphasis...	No	No	No	Yes
4. Easy to comprehend (in order of ease)...	1	2	2	3
5. Easy to compute (in order of ease).....	2	1	3	4
6. Subject to algebraic treatment.....	Yes	No	No	Yes
7. Affected by chance arrangement of data....	No	Yes	Yes	No

For example, the arithmetic mean has a rigid definition; there is no opportunity for a computer's judgment to affect its value. All items in the distribution are essential to its correct calculation. The extremely large values have most effect, while the very small items have little influence. This measure is the easiest to comprehend and is relatively easy to compute. It can be used in algebraic manipulation and is not influenced by the chance arrangement of data. In similar fashion, each of the other measures of central tendency can be quickly and precisely described.

Following this same procedure, Figure 62 provides a guide to the choice of average according to the condition of the data. It must be recalled, however, that a single answer cannot be given as to which

FIGURE 62
GUIDE TO THE SUITABILITY OF MEASURES OF CENTRAL TENDENCY
ACCORDING TO THE CONDITION OF THE DATA

CONDITION OF DATA	ARITH- METIC MEAN	MEDIAN	MODE	GEO- METRIC MEAN
Ungrouped				
Not arrayed	Yes	No	No	Yes
Arrayed	Yes	Yes	No	Yes
Grouped in a frequency distribution				
Equal class intervals.....	Yes	Yes	Yes	Yes
Unequal class intervals.....	Yes	Yes	No	Yes
Open end intervals	No	Yes	Yes	No
Containing a zero class mark.....	Yes	Yes	Yes	No

measure of central tendency should be employed under any given set of conditions. This tabulation contains some statements that are conditional and others that are absolute. For instance, the arithmetic average might be computed in a frequency distribution with open end intervals, if those intervals contained only a few items. Yet this case is marked "no" because usually the median or mode should be used. The mode is marked "no" for distributions with unequal class intervals. This means that the mode cannot be computed without regrouping. These examples indicate that the information in Figure 62 must be used along with a knowledge of the content of this and the preceding chapter. The Figure taken alone gives to the question of choice of average an air of finality, which the authors would not wish to convey.

PROBLEMS

1. "The median gives less weight to extreme items, either large or small, than the averages of calculation." Discuss.
2. a) Compute the median value of checks handled by Federal Reserve districts.*

DISTRICTS	VALUE OF CHECKS, 1937 (billions of dollars)
Boston	16.0
New York	77.9
Philadelphia	24.3
Cleveland	24.5
Richmond	13.2
Atlanta	11.7
Chicago	33.4
St. Louis	13.3
Minneapolis	5.6
Kansas City	12.1
Dallas	8.9
San Francisco	14.5
United States	255.4

* *Statistical Abstract* (1938), p. 238.

- b) Compute the extended median.
- c) Which is preferable for these data? Why?
3. a) Compute the median and the extended median of the data of Problem 9, chapter XVI, page 411.
- b) Explain the advantages or disadvantages of the median as compared with the averages of calculation as a measure of central tendency for these data.
4. a) From the following table compute the median family income.
- b) Explain fully why it would be difficult to compute the arithmetic average income.

FAMILY INCOMES IN THE UNITED STATES 1935-36 *

INCOME CLASS (dollars)		NUMBER OF FAMILIES (000 omitted)
Under	500.....	6,711
500 and under	1,000.....	11,648
1,000 and under	1,500.....	8,734
1,500 and under	2,000.....	5,186
2,000 and under	2,500.....	2,959
2,500 and under	3,000.....	1,475
3,000 and under	4,000.....	1,354
4,000 and under	5,000.....	464
5,000 and under	10,000.....	596
10,000 and under	25,000.....	260
25,000 and under	100,000.....	65
100,000 and over.....		5
Total		39,457

* *Statistical Abstract* (1938), p. 304.

- Construct an ogive of the data in Problem 4 and find the value of the median from your diagram.
- What are the objections to determining the value of the mode from an array of the data? Explain what further steps should be taken. Use data from any two columns of Problem 10, page 385, as an example.
- Find the interpolated value of the mode of the distribution of price changes in Table 80, page 408.
 - What information concerning the shape of the distribution of price changes can be obtained from comparison of the values of the arithmetic average and mode?
- Under a wages-and-hours law it is considered desirable that the number of hours of work per week should be standardized for some 300 establishments, all now operating under similar conditions except with respect to hours of work. What should be the standardized number of hours (a) if the object is to keep the total hours of work the same; (b) if the object is to change as few establishments as possible.
- From the following frequency distribution of grades in a statistics section, find the exact value of the mode. Regroup if necessary, and explain why.

GRADES IN PER CENTS	PERCENTAGE DISTRIBUTION
38-42	2
43-47	3
48-52	6
53-57	3
58-62	8
63-67	7
68-72	10
73-77	8
78-82	15
83-87	12
88-92	14
93-97	9
98-100	3
Total	100

10. Study the data given in Part A, Table 16, page 151. Discuss questions of homogeneity and bi-modality in connection with these data.
11. What are the objections to developing the subject of averages by listing the types of data with which each average should be used?

REFERENCES

- BOWLEY, ARTHUR L., *Elements of Statistics*. London: P. S. King and Son, Ltd., 1920 (fourth edition). Chapter V.
- BURGESS, ROBERT W., *Introduction to the Mathematics of Statistics*. Boston: Houghton Mifflin Co., 1927. Chapter V.
- CROXTON, FREDERICK E., and COWDEN, DUDLEY J., *Applied General Statistics*. New York: Prentice-Hall, Inc., 1939. Chapter IX.
- RIEGEL, ROBERT, *Elements of Business Statistics*. New York: D. Appleton Century Co., 1926.
- A precise statement of the applications of averages appears in chapter XI.
- YULE, G. UDNY, and KENDALL, M. G., *Introduction to the Theory of Statistics*. London: Charles Griffin and Co., Ltd., 1937. Chapter 7.

CHAPTER XVIII

MEASURES OF DISPERSION AND SKEWNESS

INTRODUCTION

IN THE last three chapters attention has been centered on the simplest methods of describing data: first, in a frequency distribution graph, and second, in terms of a single figure, a measure of central tendency. These devices are useful and very important in summarizing data but they give only a very limited description of a distribution. From the discussion of frequency distributions, it will be recalled that there are many patterns of distributions in which data may arrange themselves. Measures of central tendency indicate the values around which all the items group themselves but they give no indication of the other characteristics of the pattern. The first step in describing this pattern is to obtain a measure of the dispersion or deviation of the distribution. The second step is to obtain a measure of the shape of the pattern, a measure of asymmetry or skewness.

DISPERSION

Dispersion of data means the scatter of the items along a range of possible values. The computation of dispersion involves the measuring of the extent of this scatter. The dispersion concept is frequently employed unconsciously. For instance, in describing the weights of the men on a football team it might be said that the lightest man weighs only 140 pounds and that the heaviest weighs 216 pounds. A more complete statement of dispersion is involved when you observe that the majority of students in your statistics class are between 17 and 19 years of age, but that there is one student who must be a prodigy for he is only 15 years old, and there are two graduate students who are at least 24 or 25 years of age. Description of this kind, which gives these details about the ages of the members of a class, provides a better idea of the actual age distribution than a single statement that the average age of the members of the class is $18\frac{1}{2}$ years.

Dispersion is commonly though less precisely used in other ways. For instance, a student is reported as "better than the average," mean-

ing that his grades are higher than the average of the grades in the class. In this case there is no effort to describe the way all the grades are scattered. The statement merely points out that the grades of a given student deviate in a certain direction from the average. In golf and other games the deviations of scores above or below par are excellent criteria of the abilities of the players.

Measures of dispersion are frequently used rather loosely as norms or standards. It is not uncommon to hear a student say, "The average grade on our last exam was 81 and I made 79 so I made about the average." There is probably no great error in this type of reasoning for it is possible that no student's grade was exactly equal to the average. On the other hand, without knowing the dispersion it would be hard to say how far a particular grade might deviate from the average and still be called "about the average." It is the purpose of statistical measures of dispersion to develop standards which will provide answers to such questions.

More precise results are obtained by measuring dispersion in numerical terms and by comparison of these measures with measures of central tendency. As the student progresses in this chapter he will recognize basic similarities to methods involved in calculating measures of central tendency and will realize the need for a complete understanding and knowledge of those methods before proceeding. The criteria which were employed in evaluating the different measures of central tendency will also be used to evaluate the various measures of dispersion.

Measures of dispersion will be discussed under two major heads according to the methods involved in their computation: (1) position measures, which depend upon the values of items at certain locations in a distribution and, (2) calculated measures, which depend upon the use of the values of every item in the distribution.

Position Measures of Dispersion

The Range.—The simplest of all the measures of dispersion is the range, which is the difference between the greatest and the least values of the data. It is this measure which is in most frequent every-day use. Lists of stock prices include the high and low for the day. Weather reports state the highest and lowest temperature for the day.

When used in conjunction with a measure of central tendency the range adds greatly to one's knowledge of a distribution. For

instance, the range of the rents paid in Columbus, Ohio, as determined from the array shown in Table 57, page 352, is \$87, the difference between the lowest rent paid, \$8, and the highest, \$95. The value of the range together with the value of the arithmetic average of the rentals, \$40.69 (as computed from the ungrouped data, chapter XVI, p. 389) gives more information about the distribution than the mean alone would give. The lowest rent is about \$33 less than the average, but the highest is \$54 greater. Hence the average falls in the lower half of the range and concentration of the rents below the average can be expected.

After data have been grouped it may be impossible to ascertain the precise range of the data. In lieu of the actual end values the lower and upper limits of the lowest and highest class intervals must be used to determine the range in a frequency distribution. If the distribution in Table 88 is used, the range of rentals in Columbus, Ohio, is found to be \$90 (\$97.50 — \$7.50). This value is \$3 larger than the value obtained from the array, but similar conclusions may be drawn from it.

Partition Values.—Partition values are those which divide an array or distribution into several parts, each containing an equal number of items. They are very useful in indicating the amount of dispersion in a group of data. In employing position measures of dispersion, just as in dealing with averages of position, the items must be arranged in order of size.

Quartiles: The quartiles are the partition values which divide the items into four equal groups. That is, the first quartile, Q_1 , separates the first quarter of the total number of items from the second quarter; and the third quartile, Q_3 , separates the third quarter from the fourth quarter. Consequently the interquartile range, $Q_3 - Q_1$, includes the middle half of the items. The median, as explained in the preceding chapter, divides the first half of the items from the second half, and therefore falls between the second and third quarters. It is often referred to as Q_2 , since it is actually one of the three quartering values.

In Ungrouped Data: In an array of ungrouped data it is rather difficult to determine the values that divide the middle half of the items from the quarters at either end, unless the number of items happens to be divisible by 4. For example, an array of twelve items could be separated into four quarters of three items each, the first quartile value being halfway between the values of the third and

fourth items, and the third quartile between the values of the ninth and tenth items. For any number not divisible by 4, the "middle half" of the actual items in the sample cannot be selected, and any elementary measure¹ of intermediate values is at best a doubtful approximation. Consequently if the array consists of any considerable number of items, the best way to determine quartile values is to group the data in a frequency distribution in order to deal with item ranges instead of the actual values of the items.

In Grouped Data: The method of calculating quartile values in grouped data is shown in Table 88 for the sample of grouped rentals in Columbus. We are now dealing with item ranges; hence the item one-fourth of the distance from the lowest to the highest value is the 38.75th rental, and the item one-fourth of the way from the highest to the lowest value, or three-fourths the way from the lowest to the highest, is 3×38.75 , or the 116.25th. The values in the class intervals corresponding to these two items are the first and third quartiles respectively, and they are calculated in exactly the same way as the median. The first quartile item is the 22.75th one in the \$17.50 to \$27.50 class, since $38.75 - 16 = 22.75$.

The value of $Q_1 = \frac{(17.50 \times 4.25) + (27.50 \times 22.75)}{27} = \25.93 . The

third quartile item is the 12.25th one in the \$47.50 to \$57.50 class, since $116.25 - 104 = 12.25$.

The value of $Q_3 = \frac{(47.50 \times 5.75) + (57.50 \times 12.25)}{18} = \54.31 .²

These values of the first and third quartiles add specific information about the distribution of rentals to that which was available from the measures of central tendency. It is now known that 25 per cent of the rentals paid in Columbus are lower than \$25.93, 25 per cent are higher than \$54.31,³ and 50 per cent of the rentals fall between \$25.93 and \$54.31. The middle half of the rentals paid are therefore within a range of \$28.38.

¹ For formulas designed to provide for these cases, see Arthur L. Bowley, *Elements of Statistics* (4th ed., London: P. S. King & Son, Ltd., 1920), p. 107.

² These computations are based on the formula for the median given in chapter XVII, p. 417, with the obvious change from the median class to the classes containing the respective quartiles.

³ This is a numerical verification of the approximate results obtained from the ogive in chapter XV, Figure 54, page 373.

In the preceding chapter, the median was calculated to be \$35.34. With this information, it can be seen that the second 25 per cent of the rentals fall between the values of \$25.93 and the median \$35.34,

TABLE 88
COMPUTATION OF QUANTILES FOR THE FREQUENCY DISTRIBUTION
OF RENTALS PAID BY 155 FAMILIES IN COLUMBUS, OHIO

CLASS INTERVAL	FREQUENCY	FREQUENCIES CUMULATED TO REACH CLASS CONTAINING	
		Q_1	Q_3
\$ 7.50 and under \$17.50.....	16	16	16
17.50 and under 27.50.....	27	..	43
27.50 and under 37.50.....	44	..	87
37.50 and under 47.50.....	17	..	104
47.50 and under 57.50.....	18	..	..
57.50 and under 67.50.....	11	..	..
67.50 and under 77.50.....	10	..	..
77.50 and under 87.50.....	9	..	..
87.50 and under 97.50.....	3	..	..
Total	155	..	..

that is, between Q_1 and Q_2 , while the third 25 per cent of them lie between \$35.34 and \$54.31, or between Q_2 and Q_3 . The greatest concentration of the rentals, therefore, is in the second quarter of the distribution where 25 per cent of the rentals fall within a range of \$9.41, whereas in the third quarter of the distribution 25 per cent of the rentals fall within a range of \$18.97. The concentration in the first quarter is about the same as the third, but in the fourth quarter the rentals are dispersed over a much wider range of values.

The interquartile range ($Q_3 - Q_1$) includes the central half of the items. It is the range of values in the distribution within which there is an equal likelihood, or a 50-50 chance, that any rental selected at random in Columbus will fall. The expression "quartile deviation" usually refers to one-half the interquartile range, or $\frac{Q_3 - Q_1}{2}$, which is also called the semi-interquartile range. The value of the semi-interquartile range in this case is $\frac{\$28.38}{2} = \14.19 .

If the distribution of data were perfectly symmetrical, the semi-interquartile range would be equal to the range between the first quartile and the median or between the median and the third quartile, and would measure the range on either side of the median within

which one-fourth of the items fall. Even though actual distributions are seldom normal, the interquartile range is a useful measure so long as the departure from normal is not too marked. In practice, the semi-interquartile range is preferable to the interquartile range because it is comparable with other measures of dispersion.

Other partition values: There is no reason for using quartiles to measure dispersion in preference to any other partition values of a distribution except custom and the ease of comprehending quarter parts and the middle half. In fact, there are several other measures which are widely used, such as quintiles, deciles, and percentiles. Any of these partition values is obtained by a computation parallel to those used for the median and quartiles.

Partition measures of dispersion, like position measures of central tendency, ignore some of the items in a distribution and depend entirely upon the values which happen to attach to the items at the *points* in the distribution which are selected.

Calculated Measures of Dispersion

Average Deviation.—The average deviation employs all data in a distribution and measures how much *on the average* the items deviate from some measure of central tendency—usually the arithmetic mean or the median. The concept is so simple that its name indicates how the average deviation is calculated.

From ungrouped data: The average deviation of the examination grades from Table 70, page 388, is calculated in Table 89. First, the difference is obtained between each grade and the average grade; the sum of these differences is then divided by the number of differences (which is the same as the number of items) and the quotient, 4.96,

TABLE 89
COMPUTATION OF AVERAGE DEVIATION OF EXAMINATION GRADES

	GRADE	DEVIATION FROM AVERAGE
First	75	12.2
Second	87	0.2
Third	88	0.8
Fourth	93	5.8
Fifth	93	5.8
Total	436	24.8
Average	87.2	4.96

$$\text{Average Deviation} = \frac{24.8}{5} = 4.96$$

is the average of the deviations. It must be observed that the direction of the differences is disregarded, i.e., the differences are added without regard to algebraic sign. If the signs of the deviations from the average were regarded, the sum of the deviations, by the very nature of the arithmetic average, would be zero.

The average deviation of approximately 5 grade points indicates that this student's work was consistent. If, for instance, the arithmetic average of another student's work were likewise 87, but with an average deviation of 15 grade points, much larger variation in the grades would have been reflected, indicating much greater inconsistency in the work of the second student.

From grouped data, direct method: The average deviation for ungrouped data can be calculated without any special arrangement of the data. When there are a great many items, however, and it is necessary to group them in a frequency distribution, the average deviation

TABLE 90
COMPUTATION OF AVERAGE DEVIATION OF RENTALS PAID
BY 155 FAMILIES IN COLUMBUS, OHIO

CLASS INTERVAL	(1) MID- POINT	(2) FRE- QUENCY	(3) (4) DIRECT METHOD		(5) (6) (7) SHORT METHOD		
			Deviation from Mean (\$40.89)	Deviation (2) × (3)	<i>f</i>	<i>d_s</i>	<i>fd_s</i>
\$ 7.50–\$17.50	\$12.50	16	\$28.39	\$454.24	..	..	..
17.50– 27.50	22.50	27	18.39	496.53	..	..	..
27.50– 37.50 ...	32.50	44	8.39	369.16	..	..	..
37.50– 47.50 ...	42.50	17	1.61	27.37	17	0	0
47.50– 57.50 ...	52.50	18	11.61	208.98	18	1	18
57.50– 67.50 ...	62.50	11	21.61	237.71	11	2	22
67.50– 77.50	72.50	10	31.61	316.10	10	3	30
77.50– 87.50 ...	82.50	9	41.61	374.49	9	4	36
87.50– 97.50 ...	92.50	3	51.61	154.83	3	5	15
Total		155		\$2,639.41	68	..	121

Direct Method: Average Deviation = $\frac{\$2,639.41}{155} = \17.03

Short Method: Average Deviation = $\frac{2}{155} [(121 \times 10) + (42.50 - 40.89) 68]$
 $= \frac{2}{155} [1210 + 109.48] = \frac{1319.48}{77.5} = \17.03

can be computed as illustrated in Table 90. On the average, each rental in the sample deviates from the arithmetic average rental by \$17.03. The average deviation is \$2.84 greater than the semi-inter-

quartile range, \$14.19. The range of the arithmetic mean plus and minus the average deviation is therefore \$5.68 larger than the inter-quartile range; consequently it contains more than 50 per cent of the items, which was the proportion of rentals included in the inter-quartile range.

In the illustrations of the calculation of the average deviation the arithmetic average has been used as the measure of central tendency from which to take the deviations. The median can also be employed for this purpose. In fact, the average deviation is a minimum when measured from the median,⁴ and is therefore at its best as a measure of dispersion when the median is employed as the measure of central tendency. If the distribution of data is symmetrical, it is immaterial which of these two measures of central tendency is used, for they coincide.

An example of the use of the average deviation is furnished in the study, *Rural Regions of the United States* by Dr. A. R. Mangus. The average deviation above and below the median was used to establish limits within a sample of rural counties by which "plane-of-living" indexes might be obtained.

In the Northeastern Great Plains Region, for example, the median plane-of-living index for 1930 was 113 and the counties in the region scattered above and below that median with an average deviation of 17 units. The subtraction of the average deviation from the median gave an index of 96 while the addition of the average deviation gave an index of 130. If consideration is given to this factor alone, any three counties having plane-of-living indexes approximating 96, 113, and 130, respectively, would be representative of the region with respect to the median and average deviation of this factor.⁵

The average deviation has not been widely used, except for special purposes, largely because it is not convenient for algebraic treatment.

From grouped data, short method: When the deviations are measured from the arithmetic average⁶ and the class intervals are equal in width, the average deviation can be computed by a shorter method.⁷ This method is illustrated for Columbus rentals on the right side of Table 90. The steps in the process are:

⁴ For proof, see G. Udny Yule and M. G. Kendall, *An Introduction to the Theory of Statistics* (London: Charles Griffin and Co., Ltd., 1937), p. 145.

⁵ A. R. Mangus, *Rural Regions of the United States*, Works Progress Administration (Washington: United States Government Printing Office, 1940), pp. 95, 96.

⁶ The method is not applicable when deviations are measured from the median.

⁷ The genesis of this method can be found in John F. Kenney, *Mathematics of Statistics* (New York: D. Van Nostrand Co., Inc., 1939), p. 83.

- 1. As an origin assume the midpoint of the class containing the arithmetic average.
- 2. Write the step deviations of classes greater than the assumed origin, column 6.
- 3. Multiply the step deviations by the corresponding frequencies, and sum the products, column 7.
- 4. Multiply this sum by the width of the class interval.
- 5. Subtract the true average from the assumed average and multiply the result by the total frequency in classes with midpoints exceeding the true average, sum of column 5.⁸
- 6. Add the results of steps 4 and 5, taking account of the algebraic sign of step 5.
- 7. Multiply this sum by 2 and divide by the total frequency.

There is some question of the practical value of any short-cut method of computing the average deviation dispersion. Nevertheless, this method will produce considerable saving of time for a student who masters the several steps.

Standard Deviation.—The standard deviation, also called the root-mean-square deviation, is defined as the square root of the mean of the squared deviations. The deviations of the items from their arithmetic average are squared, these squares are summed, the sum is divided by the number of items, and the square root of the result is extracted. Prepared tables of squares (see Appendix D, or Barlow's Tables), greatly reduce the amount of work involved in obtaining the squares of the deviations. The standard deviation is usually designated either by the small Greek letter sigma, σ , or by S.D.

TABLE 91
COMPUTATION OF STANDARD DEVIATION OF EXAMINATION GRADES

GRADE X	DEVIATION FROM MEAN $(X - M)$	DEVIATION SQUARED $(X - M)^2$
75	— 12 2	148 84
87	— 0 2	.04
88	0 8	.64
93	5 8	33.64
93	5 8	33.64
436 $\div$ 5 = 87.2 = M		216.80

$$\sigma = \sqrt{\frac{216.8}{5}} = \sqrt{43.36} = 6.58$$

⁸Note that the frequency of the *average* class, whose deviation is zero, will be included in this sum whenever the true average is less than the midpoint; it will be excluded when the average exceeds the midpoint.

TABLE 92
COMPUTATION OF STANDARD DEVIATION OF EXAMINATION GRADES
BY SHORT METHOD

Grade	X	X ²	
75	.	5,625	M = 87.2
87	..	7,569	M ² = 7,603.84
88		7,744	
93		8,649	$\sigma = \sqrt{\frac{38,236}{5} - 7,603.84} = \sqrt{7,647.2 - 7,603.84}$
93		8,649	$= \sqrt{43.36}$
		38,236	$= 6.58$

From ungrouped data: The calculation of the standard deviation from ungrouped data is demonstrated by direct computation in Table 91. The algebraic formula for the standard deviation is:

$$\sigma = \sqrt{\frac{\sum (X - M)^2}{N}}$$

where X stands for each individual item; M stands for the arithmetic average of X ; N stands for the number of items.

By a slight algebraic manipulation⁹ the formula becomes:

$$\sigma = \sqrt{\frac{\sum X^2}{N} - M^2}$$

The computation of grades by this formula is shown in Table 92.

The value of the standard deviation (6.58) of the grades is approximately one and one-half grade points larger than the average deviation (4.96). The standard deviation will always be larger than the average deviation because the squaring of the deviations puts more emphasis upon extreme items in the distribution.

From grouped data, direct method: An illustration of the calculation of the standard deviation from grouped data is provided in Table 93, using the distribution of rentals of 155 families in Colum-

⁹ This formula can be developed from the definition of the standard deviation as follows:

$$\begin{aligned}
 \sigma &= \sqrt{\frac{\sum (X - M)^2}{N}} \\
 &= \sqrt{\frac{\sum X^2}{N} - \frac{2\sum (XM)}{N} + \frac{NM^2}{N}} \\
 &= \sqrt{\frac{\sum X^2}{N} - 2M \times M + M^2} \quad \text{since } \frac{\sum X}{N} = M \\
 &= \sqrt{\frac{\sum X^2}{N} - M^2}
 \end{aligned}$$

bus, Ohio. The standard deviation of the rentals paid amounts to \$20.68 as compared with the average deviation, \$17.03. It should be

TABLE 93
COMPUTATION OF STANDARD DEVIATION OF RENTALS
PAID BY 155 FAMILIFS IN COLUMBUS, OHIO

CLASS INTERVAL	(1) MID- POINT X	(2) FRE- QUENCY f	(3) DEVIATION FROM MEAN (\$10.89) (X - M)	(4) DEVIATION SQUARED (X - M) ²	(5) f (X - M) ²
\$ 7.50-\$17.50	12 50	16	- 28 39	805.9921	12895.8736
17.50- 27.50	22 50	27	- 18 39	338.1921	9131.1867
27.50- 37 50	32 50	44	- 8 39	70 3921	3097.2524
37.50- 47.50	42.50	17	1.61	2 5921	44 0657
47 50- 57.50	52 50	18	11 61	134 7921	2426.2578
57.50- 67.50	62 50	11	21.61	466.9921	5136 9131
67.50- 77.50	72.50	10	31.61	999.1921	9991.9210
77 50- 87 50	82 50	9	41.61	1731 3921	15582.5289
87.50- 97.50	92 50	3	51 61	2663.5921	7990.7763
Total		155			66296.7755

$$\sigma = \sqrt{\frac{66296.7755}{155}} = \sqrt{427.7211} = \$20.68$$

observed that both of these measures of dispersion are expressed in the same units as the data. Rentals are expressed in dollars and so are the average deviation and the standard deviation. The standard deviation is larger than the average deviation by approximately 18 per cent.

TABLE 94
SHORT METHOD OF COMPUTING STANDARD DEVIATION OF RENTALS PAID
BY 155 FAMILIES IN COLUMBUS, OHIO

CLASS INTERVAL	MIDPOINT X	FRI- QUENCY f	MIDPOINT SQUARED X ²	fX ²
\$ 7.50 and less than \$17.50 ..	12 50	16	156 25	2500 00
17 50 and less than 27.50 . .	22 50	27	506 25	13668 75
27 50 and less than 37 50 . . .	32 50	44	1056 25	46475 00
37 50 and less than 47 50 . . .	42 50	17	1806 25	30706 25
47 50 and less than 57.50 . . .	52 50	18	2756 25	49612 50
57 50 and less than 67 50 . . .	62 50	11	3906 25	42968 75
67 50 and less than 77 50 . . .	72 50	10	5256 25	52562.50
77 50 and less than 87.50	82 50	9	6806 25	61256 25
87.50 and less than 97.50	92 50	3	8556 25	25668 75
Total		155		325418.75

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum (fX^2)}{\sum f} - M^2} = \sqrt{\frac{325418.75}{155} - (40.89)^2} \\ &= \sqrt{2099.4758 - 1671.9921} \\ &= \sqrt{427.4837} \\ &= \$20.68\end{aligned}$$

From grouped data, short method: The standard deviation of grouped data can be computed by the short formula demonstrated for ungrouped data in Table 92. The work is carried out for Columbus rentals in Table 94. The result is identical with that obtained in Table 93.

Both of these computations require too much labor. A shorter method is ordinarily used to compute the value of the standard deviation. It depends upon the use of an assumed average and step deviations similar to the step method developed in obtaining an arithmetic average.¹⁰ The detailed calculations are shown in Table 95 for the rent distribution. If the procedure at the bottom of the table is put in symbols, the formula¹¹ for the standard deviation by the short-cut method becomes:

$$\sigma = i \sqrt{\frac{\Sigma (fd_s^2)}{\Sigma f} - \left[\frac{\Sigma (fd_s)}{\Sigma f} \right]^2}$$

¹⁰ The direct computation, Table 93, and the computation using the original X 's, Table 94, both require the use of the exact value of the arithmetic average; therefore, unless the average happens not to be approximate, both methods contain errors related to significant figures. The last column of Table 93 and the computation of the square root at the bottom of Table 94 are carried to more significant figures than is warranted according to the rules of chapter II. It is proper to retain the additional figures through the intermediate stages of calculation and round off to the proper number of significant figures at the final step, but the additional figures retained may not agree in the two methods of computation. All of this complication is avoided by the use of the short-cut method presented in Table 95, because it does not employ the exact value of the arithmetic average, hence is free from the effects of rounding off the figures in the average.

¹¹ The proof of this formula follows:

$$\sigma = \sqrt{\frac{\Sigma [f(X - M)^2]}{\Sigma f}}$$

$$\Sigma f \times \sigma^2 = \Sigma [f(X - M)^2]$$

Let M' = the assumed average from which deviations are measured. Add and subtract this term in the square at the right

$$\begin{aligned} \Sigma f \times \sigma^2 &= \Sigma [f(X - M' - M + M')^2] \\ &= \Sigma [f\{(X - M') - (M - M')\}^2] \end{aligned}$$

Let $(X - M') = d$ the deviations of the midpoints of class intervals from the assumed average

Let $(M - M') = k$ the difference between the true average and the assumed average then,

$$\begin{aligned} \Sigma f \times \sigma^2 &= \Sigma [f(d - k)^2] \\ &= \Sigma [f(d^2 - 2dk + k^2)] \\ &= \Sigma (fd^2) - 2k \Sigma (fd) + k^2 \Sigma f \\ \sigma^2 &= \frac{\Sigma (fd^2)}{\Sigma f} - \frac{2k \Sigma (fd)}{\Sigma f} + \frac{k^2 \Sigma f}{\Sigma f} \end{aligned}$$

$$\text{but } \frac{\Sigma (fd)}{\Sigma f} = k$$

in which i is the width of the class intervals and d_s denotes deviations in steps from an assumed average. The larger the frequency distribution, either in number of frequencies or class intervals, and the higher the class interval values, the greater the advantage in using this short-cut method.

TABLE 95
STEP METHOD OF COMPUTATION OF STANDARD DEVIATION OF RENTALS
PAID BY 155 FAMILIES IN COLUMBUS, OHIO

CLASS INTERVAL $i = \$10$	(1) MID- POINT X	(2) FRE- QUENCY f	(3) DEVIATION FROM ASSUMED MEAN IN STEPS d_s	(4) fd_s	(5) fd_s^2 (4) $\times$ (3)
\$ 7.50 and less than \$17.50..	12.50	16	-3	-48	144
17.50 and less than 27.50 ..	22.50	27	-2	-54	108
27.50 and less than 37.50 ..	32.50	44	-1	-44	44
37.50 and less than 47.50	42.50	17	0	0	0
47.50 and less than 57.50 ..	52.50	18	1	18	18
57.50 and less than 67.50 .	62.50	11	2	22	44
67.50 and less than 77.50 .	72.50	10	3	30	90
77.50 and less than 87.50	82.50	9	4	36	144
87.50 and less than 97.50 ..	92.50	3	5	15	75
Totals		155		-25	667

$$\sigma = 10 \sqrt{\frac{667}{155} - \left(-\frac{25}{155}\right)^2} = 10 \sqrt{4.3032 - 0.0260}$$

$$\sigma = 10 \sqrt{4.2772} = \$20.68$$

Relation Between Measures of Dispersion in a Normal Distribution

The quartile deviation, average deviation, and standard deviation have a definite relationship in a normal distribution. The quartile deviation is two-thirds as great as the standard deviation and the average deviation is 80 per cent as great as the standard deviation.

i.e., the sum of the frequencies times the deviations from the assumed average divided by the total frequency gives the correction which must be added to or subtracted from the assumed average to obtain the true average. (See chapter XVI, p. 402)

hence

$$\sigma^2 = \frac{\Sigma(fd_s^2)}{\Sigma f} - 2k \times k + k^2$$

$$\sigma^2 = \frac{\Sigma(fd_s^2)}{\Sigma f} - k^2$$

$$\sigma = \sqrt{\frac{\Sigma(fd_s^2)}{\Sigma f} - k^2} = \sqrt{\frac{\Sigma(fd_s^2)}{\Sigma f} - \left[\frac{\Sigma(fd_s)}{\Sigma f}\right]^2}$$

If step deviations are used, the formula becomes

$$\sigma = i \sqrt{\frac{\Sigma(fd_s^2)}{\Sigma f} - \left[\frac{\Sigma(fd_s)}{\Sigma f}\right]^2}$$

These comparisons aid in understanding where each measure places the emphasis. Thus the quartile deviation emphasizes the items near the center of the distribution; the standard deviation is largest because it gives more weight to the deviations of extreme items; the average deviation stands in an intermediate position, giving weight to each item according to the amount of its deviation.

Actual Distribution Compared with the Normal.—The relation of the three measures in a normal distribution serves also as a standard for judging the extent to which a given distribution departs from the normal form. The measures of dispersion of Columbus rentals are compared with the normal distribution as a standard in Figure 63.

FIGURE 63
COMPARISON OF COLUMBUS RENTALS WITH NORMAL DISTRIBUTION
ACCORDING TO MEASURES OF DISPERSION

MEASURE	NORMAL DISTRIBUTION		COLUMBUS RENTALS				
	Per-centage of σ	Percentage of Items Included	Value of Measure	Per-centage of σ	Percentage of Items Included		
					Total	Less than Average	Greater than Average
$M \pm \frac{Q_3 - Q_1}{2} \dots\dots$	67	50	\$40.89 $\pm$ \$14.19	69	50	34	16
$M \pm A.D. \dots\dots$	80	58	\$40.89 $\pm$ \$17.03	82	58	39	19
$M \pm \sigma \dots\dots\dots$	100	68	\$40.89 $\pm$ \$20.68	100	67	45	22

The standard relations of the normal distribution are given at the left and the corresponding results for Columbus rentals at the right. For the normal distribution the measures are taken from the center, the point at which the mode, median, and arithmetic average coincide. The results will be the same from whichever measure of central tendency the dispersion is measured. In an actual distribution, however, the three measures usually will not coincide, hence the comparison of the actual with the normal will depend upon which measure is considered to be at the "center" of the actual distribution. To conform to subsequent use of the normal distribution in curve fitting, the arithmetic average will be taken as the center or reference value. Accordingly in Figure 63 the dispersion of Columbus rentals has been measured from the average rental of \$40.89.

In a normal distribution the quartile deviation is two-thirds as great as the standard deviation, and its range above and below the center

of the distribution includes 50 per cent of the frequencies. In the rental distribution the quartile deviation is 69 per cent of the standard deviation, and 50 per cent of the frequencies are included within the quartile range about the average. This means that the middle half of the rents is less concentrated than would be expected in a normal distribution, but that the quartile range around the average includes the expected 50 per cent of them.

The value of the average deviation in a normal distribution is 80 per cent as great as the value of the standard deviation, and a range of one average deviation on either side of the center ordinate will include 58 per cent of the frequencies. In the rental distribution the average deviation is 82 per cent as great as the standard deviation and 58 per cent of the frequencies are included within its range around the average. Hence the value of the average deviation is a little greater than would be found in a normal distribution, and a range of one average deviation on either side of the average includes about the number of frequencies that would be expected if the distribution were normal.

A range of one standard deviation on either side of the center of a normal distribution includes 68 per cent of the frequencies, and a range of one standard deviation on either side of the average of Columbus rentals includes 67 per cent of the frequencies, a nominal difference.

These comparisons appear to indicate some departure of the rental distribution from the normal form, but they do not tell the full story of the amount of asymmetry present. The situation thus far is that the grouping of the rentals within the range of the several measures of dispersion shows no marked difference from the normal form.

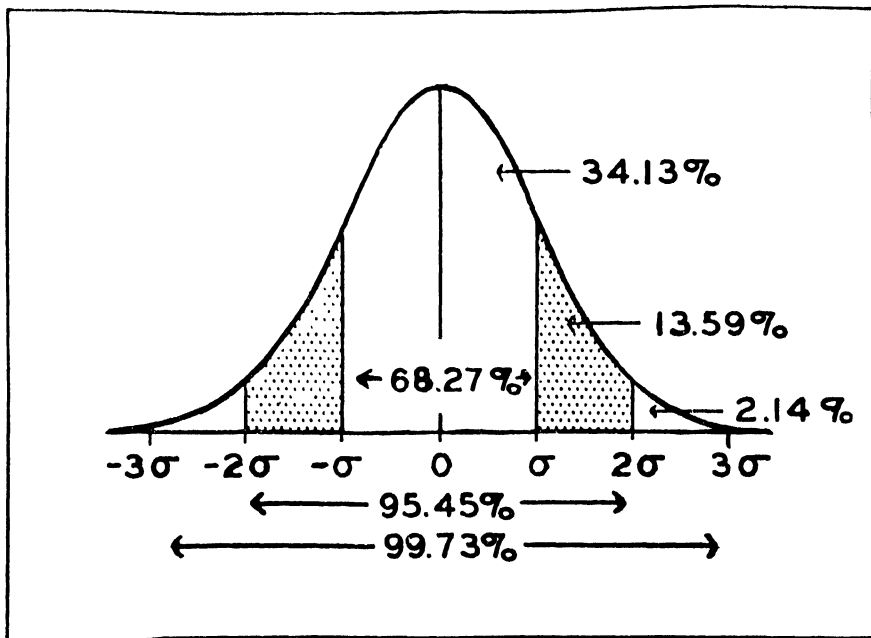
The next step in the analysis is to compare the variability on the two sides of the center. In a normal distribution no separation is necessary since the items are arranged symmetrically around the average. The columns at the right of Figure 63 show what percentage of the rentals fall below and above the average within one unit of dispersion for each of the three measures listed. A comparison of these percentages provides considerable additional information concerning the extent to which the rental distribution departs from normal form. In each measure of dispersion about two-thirds of the included items are between the average and the negative limit of one unit of dispersion, indicating a marked concentration of frequencies below the aver-

age. An exact measure of the lack of symmetry, or skewness, will be presented later in the chapter.

Importance of Standard Deviation.—In the comparison of the three measures of dispersion the standard deviation was used as the base to which the other measures were compared. The relation of the standard deviation itself to the normal distribution will now be explored a little more carefully. By methods explained in chapter XXVIII it can be shown that a range of one standard deviation above and below the center includes 68.27 per cent of the items; a range of two standard deviations above and below the center includes 95.45 per cent of the items; and a range of three standard deviations includes 99.73 per cent of the items.

These relations are shown graphically in Figure 64. The unshaded middle section delineates the 68 per cent of the frequencies falling within a range of $\pm \sigma$. The shaded portions depict the frequencies in the range between $\pm \sigma$ and $\pm 2\sigma$. The frequencies between $\pm 2\sigma$, 95 per cent of the total, are more commonly used than the shaded

FIGURE 64
FRACTIONS OF THE AREA OF THE NORMAL CURVE MEASURED BY THE STANDARD DEVIATION



27 per cent. A range of $\pm 3\sigma$ includes nearly all of the frequencies. This property of the normal distribution is widely used in analyzing actual distributions. So long as the departure of the actual from symmetry is only moderate, a range of 3σ on either side of the average will give the practical limits of the distribution.

Measures of Relative Dispersion

One might suppose that two distributions of similar data could be compared directly by means of the same measure of dispersion computed for each. But such comparisons tacitly assume the equivalence of all of the essential features of the several distributions, an assumption which is seldom warranted. It is therefore safer to use *relative* dispersion in comparing two or more distributions, even when they are measured in the same absolute units. This is always true except where the items of the two series, expressed in comparable units, are of about the same average size. For instance, the range of prices of automobiles might be \$500 and the range of prices of packaged cereals might be \$0.03. On the other hand, the absolute amounts of dispersion in two different distributions may appear to be numbers of the same size but the units of the two may be very different and not comparable. The deviation in the production of coal in the United States is expressed in tons, while the dispersion in the production of natural gas is expressed in cubic feet; tons and cubic feet cannot be compared directly. For two entirely dissimilar distributions, expressed in different units, the only possible method of comparing the degree of dispersion is by a relative measure.

Relative dispersion is the ratio of a measure of deviation to a measure of central tendency. It is called the *coefficient of variation*, and is usually denoted by V . Being ratios, these relative figures are not expressed in the units of the data; they are simply abstract numbers which express the amount of any of the measures of dispersion as a proportion of a measure of central tendency, usually as a per cent.

The standard used in computing relative dispersion should be the same one that was used in measuring the deviations in the calculated measure. This means the arithmetic average for the standard deviation and either the arithmetic average or the median for the average deviation. The median can be used in changing the quartile deviation to a relative form but half the sum of the first and third quartiles

is more generally applicable and is equal to the median in a normal distribution.

The formulas of the coefficients of variation using the different measures of dispersion are as follows:

$$V_{\sigma} = \frac{\sigma}{M} \times 100$$

$$V_{A.D.} = \frac{A.D.}{M_e \text{ or } M} \times 100$$

$$V_Q = \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} \times 100 = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$$

The added interpretation gained by using coefficients instead of absolute measures can be demonstrated by comparing the dispersion of Columbus rentals with the similar distribution of Buffalo rentals that was shown in Figure 56, page 378. The various absolute measures for the two distributions are as follows:

MEASURE	COLUMBUS	BUFFALO
Range	\$ 7.50 to \$97.50	\$ 7.50 to \$97.50
Quartiles 1 and 3	25.93 and 54.31	19.42 and 36.56
Arithmetic average	40.89	29.62
Average deviation (from M)	17.03	11.12
Standard deviation	20.68	14.37

Thus, although the two distributions are identical in range, the average is much lower for Buffalo and the dispersion is much smaller. When relative measures of dispersion are computed, however, the difference between the two cities is decreased considerably. This is, of course, due to the use of the smaller denominators for the Buffalo data.

RELATIVE MEASURE	COLUMBUS	BUFFALO
$\frac{\sigma}{M}$	$\frac{20.68}{40.89} = 50.6\%$	$\frac{14.37}{29.62} = 48.5\%$
$\frac{A.D.}{M}$	$\frac{17.03}{40.89} = 41.4\%$	$\frac{11.12}{29.62} = 37.5\%$
$\frac{Q_3 - Q_1}{Q_3 + Q_1}$	$\frac{54.31 - 25.93}{54.31 + 25.93} = 35.4\%$	$\frac{36.56 - 19.42}{36.56 + 19.42} = 30.6\%$

Criteria of Measures of Dispersion

As was mentioned earlier in this chapter, measures of dispersion can be tested by the same criteria as were used for evaluating the

measures of central tendency. The different measures of dispersion are listed in Figure 65 according to the way they satisfy the criteria.

FIGURE 65
SUMMARY OF CRITERIA OF MEASURES OF DISPERSION

CRITERION	RANGE	PARTITION VALUES (Quartiles, Percentiles, Deciles, etc.)	AVERAGE DEVIATION	STANDARD DEVIATION
1. Rigid definition	No	No	Yes	Yes
2. All items are necessary	No	No	Yes	Yes
3. Affected by value of extreme items....	Yes	No	Yes	Yes
4. Easy to comprehend (in order of ease) .	1	2	3	4
5. Easy to compute (in order of ease) ...	1	2	3	4
6. Subject to algebraic treatment.....	No	No	No	Yes
7. Affected by chance arrangement of data.	No	Yes	No	No

In Figure 65 there are two kinds of marks to indicate the relations of the various measures to each of the criteria. A mark of "yes" or "no" indicates that the criterion is or is not satisfied by the measure in question. For instance, extreme items affect the range, the average deviation and the standard deviation; they do not affect the partition measures of dispersion. The numbers 1, 2, 3, 4, in the order named, show the extent to which the measures meet the criteria. Opposite criterion number 5, for instance, number 1 indicates that the range is the easiest of the four measures to calculate.

Examination of the list shows that the average deviation and standard deviation seem to satisfy the criteria about equally well. The average deviation is rated as easier to comprehend and calculate than the standard deviation, but the standard deviation is susceptible to algebraic manipulation, and the average deviation is not. This last is a very important characteristic of a measure of dispersion since very frequently the dispersion of a distribution must be used in conjunction with other statistical measures and in algebraic development. In addition, the standard deviation has recognized relationships to the normal curve which are not shared by the other measures of dispersion. The standard deviation consequently is used much more widely than any other measure of dispersion.

SKEWNESS

The skewness of a distribution is the extent to which it is distorted, or the degree to which it deviates from symmetry. The values of the

arithmetic mean, the median, and the mode are identical in a symmetrical distribution of data. If the values of the three measures are not the same, the distribution is not symmetrical. It is only necessary to study the general relationship between the mean, median, and mode to develop rough indicators of the lack of symmetry.

If a distribution has in it a few very high-valued items, the arithmetic average will be affected by these values, whereas the median is only affected by the number of such items, and the mode is not affected at all by their presence. Hence the mean will exceed the median and the median will exceed the mode. When this is the situation, the distribution is said to be skewed to the right. If, on the other hand, there are a few very widely dispersed low items, distortion of the curve is in the direction of the lower values or to the left, and the mean will be smaller than the median and the median will be smaller than the mode. (See Figure 57, C and D, page 380.)

But the amount of difference between the averages offers no basis for judging the meaning of the skewness until some reference base is established. The natural base is a measure of dispersion, i e., the lack of symmetry can be referred to the lack of uniformity. The standard deviation is used for this purpose because it lends itself to algebraic treatment.

Method of Averages.—In conformity with this principle, relative skewness has been defined by Karl Pearson as follows:

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \frac{M - Mo}{\sigma}$$

The solution of this formula provides an indicator of both direction and extent of skewness. If the skewness is toward the higher values, the mean is larger than the mode and the value of the skewness is positive. When distortion occurs toward the lower values, the mode is larger than the mean and the skewness is negative. The value of skewness is zero when the distribution is symmetrical, for then the mean and mode are equal. The greater the lack of symmetry, the higher will be the value obtained by solving this formula.

The relative skewness of the distribution of rents in Columbus would be computed as follows:

$$M = \$40.89; Mo = \$31.99; \sigma = \$20.68$$

$$\text{Skewness} = \frac{40.89 - 31.99}{20.68} = + \frac{8.90}{20.68} = + 130 \text{ or } + 43.0 \text{ per cent}$$

This result confirms in figures the moderate positive skewness that could be seen in figure 52, page 367, and in the range and quartile measures of dispersion in this chapter.

When there is only a moderate degree of skewness, i.e., when the difference between the arithmetic average and the mode is not greater than approximately 75 per cent of the standard deviation, the skewness may be calculated from the formula:

$$Sk = \frac{3 (\text{Mean} - \text{Median})}{\text{Standard Deviation}} = \frac{3 (M - Me)}{\sigma}$$

The use of this formula should be confined to those cases in which the mode is not well defined.

Quartile Method.—Bowley has suggested a measure of skewness¹² which can be computed from quartiles by the formula:

$$Sk = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)} = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1}$$

The values obtained from this measure may vary from 0 (for perfect symmetry) to ± 1 . The results are not comparable with those obtained from the Pearsonian formula.

USES OF MEASURES OF DISPERSION AND SKEWNESS

There are many uses of measures of dispersion in addition to those already illustrated. Opportunities for their application will become more apparent as experience with the analysis of data is gained. Students at this point in their studies of statistical measures frequently want to know what specific uses there are for the various measures they learn to compute. The following short summary indicates various uses of measures of dispersion and skewness.

Aid in Description

The simplest and most common use of a measure of dispersion is in the description of data. The averages give the values which are typical for a group of data, but the measures of dispersion provide the basis for evaluating the extent of the scatter of the data. These measures simply add more information about a distribution or series of data than would be available if only the mean, median, or mode

¹² Arthur L. Bowley, *Elements of Statistics* (4th ed., London: P. S. King and Son, Ltd., 1920), p. 116.

were known. Measures of skewness, indicating the extent and direction of the lack of symmetry of a distribution, further augment the knowledge of its characteristics.

Comparison of Dispersion

The measures of central tendency of two sets of data may be very similar, but the range and pattern of scatter or arrangement of the data may be very different. The measures of dispersion of similar data can be compared in absolute units to show whether the different data are similar in range and deviate from approximately equal typical values. Even if several sets of data are expressed in different kinds of units or in similar units but with totally different levels of value, their dispersion can be compared relatively through the coefficient of variation.

Provision of a Standard

By the use of measures of dispersion, and particularly of the standard deviation, comparison of the dispersion in a given group of data with that of the normal curve as a standard is made possible. It was pointed out on page 451 that approximately 68 per cent of all the items in a normal distribution are included within one standard deviation above and one standard deviation below the arithmetic average. With these known characteristics, the standard deviation becomes a standard unit of measure by which any group of data can be compared with a normal distribution.

Development of Other Measures

The standard deviation, the most commonly employed of the various measures of dispersion, is of wide use in the development of other statistical measures. It is important that every student should understand the calculation, construction, and characteristics of the standard deviation so that he will be able more easily to comprehend the meaning of such statistical developments as the coefficient of correlation and variance.

Evaluation of Measures of Central Tendency of a Sample

It has been pointed out in some of the illustrations of the calculation and interpretation of the measures of central tendency that the data represent samples. The measures of central tendency might be

different in other samples. A question concerning the variation of these measures in a number of samples or their reliability in a single sample is raised at once. The standard deviation is employed in evaluating this variation or reliability. The character of the use of the standard deviation for this purpose will be explained in chapter XXIX.

PROBLEMS

1. Explain the purpose of measuring (a) dispersion; (b) skewness.
2. Explain the difficulty that would be encountered in finding the values of Q_1 and Q_3 in the array of 65 wages in Table 81, page 415.
3. The following are the weekly earnings of workers in a canning factory for the months of February and September, 1937:

WEEKLY EARNINGS	No. OF WORKERS	
	FEB.	SEPT.
Less than \$ 6.00	8	1
\$ 6.00- 8.99	12	0
9.00- 11.99	14	1
12.00- 14.99	21	0
15.00- 17.99	9	22
18.00- 20.99	6	34
21.00- 23.99	0	41
24.00- 26.99	0	20
27.00- 29.99	0	5
30.00 and over	2	20
	72	144

Determine the accuracy of each of the following statements (show your computations):

- a) The maximum weekly earnings received by the lower half of the workers in February was less than the maximum received by the lowest fourth in September.
 - b) The relative range of wages was greater in September than in February.
 - c) The most commonly received wage was twice as high in September.
 - d) The dispersion was greater in September than in February.
4. Given the following frequency distribution:

INCOMES	No. OF PERSONS
\$1,000-\$1,024.9	20
1,025- 1,049.9	60
1,050- 1,074.9	120
1,075- 1,099.9	150
1,100- 1,124.9	110
1,125- 1,149.9	80
1,150- 1,174.9	10

- a) Find two incomes between which the central half of the persons is to be found.

- b) Find an income such that there are as many persons earning less as there are more.
5. From the data in Table 16, page 151, discuss the question of uniformity of earnings for different grades of skill. Compute whatever measures are pertinent to your discussion.
6. a) Why is the standard deviation dispersion always greater than the average deviation dispersion?
b) Given the monthly sales of a group of eight drugstores:

\$4,320	\$2,840
2,864	3,347
3,762	4,112
3,216	3,491

- (1) Compute the average deviation dispersion of the sales of these stores.
- (2) Compute the standard deviation dispersion.
7. A purchasing agent received samples of manila envelopes from two suppliers. He had the samples tested in his own laboratory for tearing weight, with the following results:

TEARING WEIGHT IN LBS.	SAMPLES FROM	
	Company A	Company B
50-59.9	3	10
60-69.9	42	16
70-79.9	12	26
80-89.9	3	8
	60	60

- a) Which company's envelope has the higher average resistance to tearing?
b) Which company's envelope is more uniform?
8. A manufacturer producing galvanized steel sheets coated to 2 ounces per square foot wished to establish the tolerance which he could quote in his catalog. A minimum of three standard deviations below the 2 ounce standard would protect him against more than nominal rejections of sheets by the buyer. At what maximum should the tolerance be set according to the following sample?

WEIGHT PER SQUARE FOOT (in ounces)	NUMBER OF PIECES
1.75 and less than 1.80	2
1.80 and less than 1.85	8
1.85 and less than 1.90	9
1.90 and less than 1.95	16
1.95 and less than 2.00	26
2.00 and less than 2.05	35
2.05 and less than 2.10	19
2.10 and less than 2.15	6
2.15 and less than 2.20	1

9. a) Referring to the distribution of price changes in Table 80, page 408, compute the skewness by two methods.
- b) The price changes in this distribution are skewed in the positive direction. Explain why.
- c) What general statement can be made concerning the direction and amount of skewness present in distributions of price changes?

REFERENCES

BURGESS, ROBERT W., *Introduction to the Mathematics of Statistics*. Boston: Houghton Mifflin Co., 1927.

Chapter VII contains a semi-mathematical presentation of the subject.

KENNEY, JOHN F., *Mathematics of Statistics*. New York: D. Van Nostrand Co., Inc., 1939.

Part I, Chapter V, contains a brief but complete mathematical presentation of methods of measuring dispersion. Chapter IV gives the complete notation of the method of moments.

MILLS, FREDERICK C., *Statistical Methods*. New York: Henry Holt and Co., 1938. Chapter V.

RIEGEL, ROBERT, *Elements of Business Statistics*. New York: D. Appleton-Century Co., 1927.

YULE, G. UDNY, and KENDALL, M. G., *Introduction to the Theory of Statistics*. London: Charles Griffin and Co., Ltd., 1937.

An excellent discussion of methods of dispersion is given in Chapter 8

CHAPTER XIX

INDEX NUMBERS

INTRODUCTION

MANY phases of modern business are described by the use of index numbers. Numerous agencies, both governmental and private, spend a great deal of time and large amounts of money in constructing index numbers to aid in the interpretation of movements in business and of economic activity. Certain statistical publications, notably the *Survey of Current Business* and the *Statistical Bulletin* of the Standard Statistics Company, contain many series of business data expressed in ratio¹ form as index numbers.

Uses of Index Numbers

Index numbers are the most widely used ratios in statistics. In contrast with actual data, they have the following important advantages:

1. They reduce large and cumbersome figures to a form in which they can be more widely used and more easily comprehended. The index number is usually expressed as a relative of some base figure instead of in actual monetary units such as dollars per ton, cents per bushel, etc.

2. They provide a method of measuring relative changes from time to time or from place to place. It is possible to compare 35 cents for a dozen eggs with 25 cents for a pound of bacon; but it is not so easy to compare price changes in the two articles over a period of time. Index numbers of the egg and bacon prices would indicate the relative change in each price from some given price, and which of the two prices had shown the greater change. As the number of items increases, this advantage becomes even more apparent. On May 10, 1941, the United States Bureau of Labor Statistics Index of Wholesale Prices for 887 commodities stood at 84.0. This single figure is an expression of the average relation of prices on May 10, 1941, to prices in 1926. Such indexes are not restricted to the measurement of changes through time. For instance, on a certain date the amount of factory employ-

¹ Practically all index numbers are ratios, although one of the basic forms, the aggregative index discussed on pp. 468-69, is simply a total.

ment for each city in a group of cities might be compared with that in one given city used as a standard.

3. They facilitate comparison of changes in series of data expressed in a variety of units, e.g., dollars, tons, gallons, etc.¹ Data pertaining to production, sales, inventories, costs, or other aspects of business may be put into index number form.

4. They make possible the construction of composites from series that could not otherwise be combined. Dissimilar data when expressed in relative form can be combined, and many examples of such combinations appear throughout this chapter and the next.²

5. They describe patterns of business which are due to consumption habits. Many series of monthly or weekly data reflect the occurrence of holidays and seasonal periods. The annual peak in department-store sales, for instance, regularly occurs in December; clothing sales are usually at a maximum at Easter, etc. An illustration of this use is given in detail in chapter XXIII.

6. They permit the concealment of absolute values when necessary to safeguard the identity of data. This use of index numbers has arisen out of the desire of business men to have information about business in general without revealing data of specific firms. At the present time, when much interest is centered in the collection and interpretation of business data, the problem of divulging private information to collecting agencies becomes important. A short time ago a business firm in Toledo, when asked if it would be willing to report its monthly sales in a co-operative reporting plan, responded, "Yes, but we'll only report them to you in index number form because we aren't permitted to reveal the dollar amounts." This advantage of index numbers may decrease in importance as business men come to have more faith in responsible data-collecting agencies.

Historical Development of Index Numbers

The credit for inventing index numbers has been given to an Italian, G. R. Carli, who published his work in 1764.³ Little was done in developing them, however, until the latter part of the nineteenth century. Since 1900 index numbers have become universally recognized as an

¹ For a very critical treatment of this characteristic of index numbers, see Bassett Jones, *Horses and Apples, A Study of Index Numbers* (New York: The John Day Co., 1934).

² Wesley C. Mitchell, *Index Numbers of Wholesale Prices in the United States and Foreign Countries* (Bulletin No. 284, United States Bureau of Labor Statistics, October, 1921) p. 7.

important statistical tool and have been developed in numerous forms for many uses.

KINDS OF INDEX NUMBERS

An examination of the financial section of a newspaper will reveal many different index numbers which describe changes in various aspects of business. These index numbers may be classified as: (1) price indexes, (2) quantity indexes, (3) value indexes, and (4) special purpose indexes.⁴

Price Indexes

The oldest and best known indexes are those dealing with prices. Prices have been of political interest for a long time because of their relation to the domestic economy of nations and their importance in measuring international trade.

The necessary data for price index numbers arise from the exchange of commodities, (1) at different stages of production—raw materials, semi-finished goods, and completely fabricated products; (2) at several levels of distribution—industrial, wholesale, and retail; and (3) for a variety of groups of items—consumers' goods, producers' goods, imports and exports, stocks and bonds, durable and non-durable goods. The names of a few commonly known and widely used index numbers of prices are listed in Figure 66.

Quantity Indexes

Quantity indexes are constructed from data other than prices. They may measure the volume of construction, or employment, or physical inventory, or production, or distribution. They are computed for (1) industry in general, (2) specific industries, or (3) specific operations. Such indexes may pertain to various stages of the productive or distributive process. The data may represent the country as a whole or local communities.

Because of the nature of the data, quantity index numbers are frequently less reliable than those based on dollar figures. In the past, records were designed to include chiefly those aspects of business which could be expressed in monetary units, and consequently long

⁴ In general usage the terms "index number" and "index" are used interchangeably, and either one may be applied to a series of relatives or to a single relative at a specified date. This practice will be followed in the text.

series of data in physical units are difficult to obtain. Series that have been maintained in physical units are usually so restricted in scope that they cannot be taken as representative either of all aspects of a single business or of even a single aspect of business in general. The most commonly used quantity index numbers are listed in Figure 66.

Value Indexes

Since value is the result of multiplying quantity by price, index numbers of value are measures of both quantity and price. They are used to represent total dollar amounts and usually employ a simple construction method. There are relatively few index numbers of value, but these few, some of which are listed in Figure 66, are well known.

Special Purpose Indexes

The solution of business problems, whether of an individual firm or of a whole industry, of a locality or of a nation, requires that various kinds of data be combined and measured. Index numbers can be employed for these special purposes.

In 1936 W.P.A. statisticians devised several index numbers for measurement and comparison of the effects of legislation upon the financial condition of many countries. In studying markets, analysts frequently construct indexes to indicate potential sales and sales quotas. To measure changes in general business activity, composite indexes of business activity are constructed. The components of these indexes may be either prices, quantities, or values, or several of these may be in combination. Of the many special purpose indexes, a few are listed in Figure 66.

BASIC METHODS OF CONSTRUCTING INDEX NUMBERS

The available data and the purpose of an index number generally determine the method of calculation, but the fundamentals of construction must be considered regardless of the data or the purpose.

Simple Index Numbers

A simple index number is one that is made from a single series or set of data, either internal or external, continuing over a period of time or extending from one place to another. It may be defined as an index number in which the item representing a certain period or

FIGURE 66
SOURCES OF COMMONLY USED INDEX NUMBERS

KIND	NAME OF INDEX	PREPARED BY:	PUBLISHED REGULARLY IN:
Price	Daily Weighted Index of 30 Basic Commodities	Dun & Bradstreet, Inc.	<i>Dun's Statistical Review.</i>
	Composite Price of Pig Iron	Iron Age Publishing Co.	<i>Iron Age; Survey of Current Business; Commercial and Financial Chronicle.</i>
	Industrial Commodity Prices	Babson's Statistical Organization, Inc.	<i>Babson's Reports.</i>
	Wholesale Prices	National Fertilizer Association	<i>Weekly Release of National Fertilizer Association.</i>
	Retail Price Index for Department Store Items	Fairchild Publications Corporation	<i>Daily News Record; Survey of Current Business; Commercial and Financial Chronicle.</i>
	Cost of Living Index	National Industrial Conference Board	<i>Release of National Industrial Conference Board; Survey of Current Business.</i>
	Index of Prices Paid by Farmers	United States Dept. of Agriculture	<i>Agricultural Situation; Crops and Markets.</i>
	Index of Stock Prices, Daily, Weekly, and Monthly	Standard Statistics Co. Inc.	<i>Standard Statistics Co. Bulletin.</i>
Quantity	Indexes of Factory Employment	U. S. Bureau of Labor Statistics	<i>Release of United States Bureau of Labor Statistics; Employment and Payrolls; Monthly Labor Review; Survey of Current Business; Federal Reserve Bulletin.</i>
	Index of Industrial Production	Board of Governors of the Federal Reserve System	<i>Federal Reserve Bulletin; Survey of Current Business.</i>
	Operating Rate of the Steel Industry	American Iron and Steel Institute	<i>Steel; Survey of Current Business.</i>
Value	Index of Department Store Sales	Board of Governors of Federal Reserve System	<i>Federal Reserve Bulletin; Survey of Current Business.</i>
	Index of Chain Store Sales	Chain Store Age	<i>Chain Store Age; Survey of Current Business.</i>
	Indexes of Manufacturing Payrolls	U. S. Bureau of Labor Statistics	<i>Monthly Labor Review; Employment and Payrolls; Survey of Current Business; Federal Reserve Bulletin.</i>
	Indexes of Construction Contracts Awarded	Board of Governors of Federal Reserve System (based on F. W. Dodge Corp. data)	<i>Federal Reserve Bulletin; Survey of Current Business.</i>
	Index of Number of Persons Employed in Agricultural Employment on the First of Each Month	Works Progress Administration	<i>Trends of Employment in Agriculture 1909-1936, National Research Project, W.P.A., Phila., Pa., 1938.</i>
Special Purpose	Index of Business Activity	New York Times	<i>New York Times; Survey of Current Business.</i>
	Index of Business Activity in Buffalo, N. Y.	Bureau of Business and Social Research, University of Buffalo	<i>Statistical Survey, University of Buffalo.</i>

location is written as 100 (the base) and the other items in the set are expressed as percentages of the value of the base item. This type of construction is frequently called a price relative, quantity relative, or value relative. These relatives are valuable to any organization interested in a continuous measure of change in data of general importance, or to an individual firm wishing to follow changes in data arising from its own operations.

The latter use is illustrated in the construction of a simple index of the assets of the American Sugar Refining Company⁵ as shown in Table 96. The three steps which must be followed are: (1) the choice of a base year, (2) the division of the assets in each year by the assets in the base year, and (3) the multiplication of each quotient by 100 to express it as a per cent. If 1926 is chosen as the base year, then the assets (in thousands of dollars) in every year of the series must be divided by the assets in 1926, that is, by \$161,394. The computation⁶ is performed as follows:

$$\text{for 1923 } \frac{155,779}{161,394} \times 100 = 96.52$$

$$\text{for 1926 } \frac{161,394}{161,394} \times 100 = 100.00 \quad (\text{The index number in the base period should always equal 100.})$$

$$\text{for 1936 } \frac{117,887}{161,394} \times 100 = 73.04$$

Note that in each case the quotient of the given year's assets divided by the assets in the base year was multiplied by 100. This was done in order to obtain a result expressed as a per cent, rather than as a decimal fraction. An index number is written just as a per cent, except that the per cent sign (%) is not used.

⁵ *Annual Reports of the American Sugar Refining Company.*

⁶ Whenever it is necessary to make a series of divisions which involve the use of a constant divisor, as in this instance, time can be saved when calculating machines are employed by using the reciprocal of the constant divisor as a fixed multiplier. The reciprocal of a number is 1 divided by that number. In the computation of the 1923 index number of assets above, the operation:

$$\frac{155,779}{161,394} \times 100 = 96.52$$

is the same as,

$$155,779 \times \frac{1}{161,394} \times 100;$$

or, $155,779 \times .000,006,196,02 \times 100 = 96.52.$

The series of index numbers can therefore be obtained in a single operation by multiplying each year's assets by .000,619,602. This figure can be kept in the machine without change throughout the computation, thus reducing the likelihood of error. After inserting the reciprocal on the keyboard, it should first be tested by multiplying it by the original base (to produce 100.000000 or 99.999999) before using it as a multiplier for the other years.

TABLE 96
SIMPLE INDEX NUMBER OF ASSETS OF THE AMERICAN SUGAR
REFINING COMPANY 1923-1938

YEAR	ASSETS (000 omitted)	INDEX (1926 = 100)
1923	\$155,779	96.52
1924	162,854	100.90
1925	164,064	101.65
1926	161,394	100.00
1927	158,105	97.96
1928	159,620	98.90
1929	157,128	97.36
1930	148,468	91.99
1931	139,007	86.13
1932	135,887	84.20
1933*	123,358	76.43
1934*	120,356	74.57
1935*	117,666	72.91
1936*	117,887	73.04
1937*	118,193	73.23
1938*	111,314	68.97

* Period 1923-1932 includes domestic constituent companies, but beginning in 1933, the data include domestic and Cuban constituent companies.

This simple type of construction is used when a record is desired of a single series, e.g., a commodity, brand or firm. The index may pertain to amount produced, price, amount consumed, shipped, etc. It can be computed for days, weeks, months, years, or any other desired time period. Monthly and annual index numbers are most common, and are widely used by individual businesses, research institutions, and governmental agencies. An examination of data source books will reveal many index number series of this type, of which the following examples appear regularly in the *Survey of Current Business*:

- a) Index of Foreclosures (monthly), Metropolitan Cities and Non-Farm Real Estate (1926 = 100), indicates the changes in the numbers of properties foreclosed.
- b) Index of Construction Contracts (weekly), from a daily average (weekly, 1923-25 = 100), gives a comparable basis for measuring changes in anticipated building activities.
- c) Carloadings (weekly) (1923-25 = 100) shows changes in number of freight car loads of goods transported, an indication of business activity.
- d) Commercial Failures (weekly) (1923-25 = 100) indicates the trend of the number of failures in business.

Three Composite Index Numbers

Most of the index numbers in common use are composite. They employ the principles just described for simple indexes but are de-

veloped by combining several different sets of data. In the following pages, three basic methods of constructing composite index numbers are described. These are: (1) the aggregative index; (2) the relative of aggregates; and (3) the average of relatives. The three methods will be described first without the introduction of weights, and in a later section with weights included. The name of the index describes the method of construction in each case, hence the student who understands the explanations of the methods and learns their names at the outset should not have difficulty in remembering the different procedures. Formulas for all the indexes are explained on pages 477-78.

The Aggregative Index.—As its name suggests, the aggregative index number is computed by adding together the various items of data which are to be combined. To compute an aggregative price index, for example, the items to be included are chosen and the prices of unit quantities of these items are then added. The sum, in dollars and cents, is the aggregative index.

A simple illustration of the method is shown in Table 97. The totals in row (d) in columns 2 and 3, 62.0 cents and 54.3 cents, are the aggregative index numbers of the prices of the commodities, bread,

TABLE 97
COMPUTATION OF A SIMPLE AGGREGATIVE INDEX OF PRICES OF
THREE FOOD COMMODITIES IN THE UNITED STATES FOR
THE MIDDLE OF JANUARY, 1938 AND 1939 *

COMMODITY	(1) UNIT	(2) JAN. 18 1938 (cents)	(3) JAN. 17 1939 (cents)
a) Bread, white	pound	8.9	8.1
b) Butter	pound	40.4	33.5
c) Milk, fresh delivered	quart	12.7	12.7
d) Total and aggregative index.....		62.0	54.3
e) Relative of aggregates index, row (d), cols. 2 and 3 ÷ col. 2, × 100.....		100.0	87.6

* Release of United States Department of Labor, Bureau of Labor Statistics, *Retail Food Prices by Cities*.

butter, and milk, as of the middle of January of the years 1938 and 1939. Between January, 1938, and January, 1939, the total of the three decreased 7.7 cents. This is an expression of changes in the prices of the three commodities together and does not indicate whether the price of any particular commodity has gone up or down. The changes, whether increases or decreases, in the prices of the individual items are concealed in the summation of the prices.

One of the best known and most widely used index numbers of this type in the United States is the *Dun and Bradstreet Weekly Food Price Index*. A selected portion of this index which represents the sum total of the wholesale price per pound of 31 commodities in general use is shown in Table 98.

The Relative of Aggregates Index.—The name of this index also serves as a guide to its method of construction. The aggregative index in itself affords no relative comparison. However, it can be converted

TABLE 98
DUN AND BRADSTREET'S WEEKLY FOOD PRICE INDEX FOR
CORRESPONDING EIGHT WEEK PERIODS, 1936 TO 1939 *

WEEKS	1939	1938	1937	1936
	\$	\$	\$	\$
May 30	2.25	2.34	2.85	2.54
May 23	2.25	2.35	2.85	2.55
May 16	2.27	2.35	2.86	2.52
May 9	2.28	2.34	2.84	2.54
May 2	2.27	2.36	2.82	2.58
April 25	2.27	2.36	2.86	2.60
April 18	2.28	2.37	2.89	2.59
April 11	2.28	2.37	2.89	2.64

* Dun and Bradstreet, Inc., *Dun's Review* (June, 1939), p. 39.

into relatives by expressing each aggregate as a percentage of some base period aggregate, that is, by applying to the aggregates the method of simple index number construction. The results of calculation by this method can be illustrated by using the data presented in Table 97. If the 1938 aggregate is used as the base, the relative of aggregates index number for 1939 is computed by dividing the 1939 aggregate by the base figure, 62.0 cents. In row (e) column 3, this index is shown to be 87.6. Prices on January 17, 1939, therefore, were only 87.6 per cent as high as the prices a year earlier, or, in other words, they were 12.4 per cent lower in 1939 than on the same date in 1938. Again it must be pointed out that this index number indicates only the relative change in the total prices; it does not indicate that the price of any one of the three commodities declined by this per cent.

The Average of Relatives Index.—An average of relatives price index for a given period is the arithmetic⁷ average of the ratios obtained by dividing the price of each item included in the index by the

⁷ The use of other averages in constructing this type of index will be discussed in a later section.

price of that item in the period chosen as the base. The first step in calculating this index is the division of the price of each article at every period for which an index number is desired by the price of the same article in the base period. The second step is to average these resulting price ratios.

TABLE 99
COMPUTATION OF INDEX OF RETAIL PRICES OF BITUMINOUS COAL
IN COLUMBUS, OHIO, BY AVERAGE OF RELATIVES METHOD,
SEPTEMBER AND OCTOBER, 1938

SIZE OF BITUMINOUS COAL	(1) PRICES PER TON IN SEPTEMBER 1938	(2) RELATIVES SEPTEMBER 1938 $(1) \div (1),$ $\times 100$	(3) PRICES PER TON IN OCTOBER 1938	(4) RELATIVES OCTOBER 1938 $(3) \div (1),$ $\times 100$
a) Lump	\$6.40	100.0	\$6.59	103.0
b) Egg	6.02	100.0	6.04	100.3
c) Stoker	5.39	100.0	5.89	109.3
d) Total	\$17.81	300.0	\$18.52	312.6
e) Average of relatives index, row (d), cols. 2 and 4 $\div 3$		100.0		104.2
f) Relative of aggregates index, row (d), cols. 1 and 3 $\div$ col. 1, $\times 100$		100.0		104.0

The method of computing this index is illustrated in Table 99. The index for any given period is obtained as in column 4. For each of the sizes of bituminous coal, a ratio is computed of the price in October, 1938, to the price in the base period, September, 1938. The ratio for lump coal is obtained by dividing the October price by the September base price and multiplying the quotient by 100, $\frac{6.59}{6.40} \times 100 = 103.0$. This relative figure is a simple index of the October price of lump coal, on September as a base. The October index of prices for the three sizes of bituminous coal taken together, obtained by computing the arithmetic average of the simple indexes for each of the three, $\frac{312.6}{3} = 104.2$, is shown in row (e) of column 4. This index indicates that retail prices for lump, egg, and stoker sizes of bituminous coal taken together increased 4.2 per cent over the prices of the same sizes of coal in the base period.

Table 99 also provides the basis for a review of results obtained by the other two methods. The aggregative indexes, the totals in row (d) of columns 1 and 3, show that the total cost of a ton of each of the three sizes was \$18.52 in October, 1938, as compared with \$17.81 in September, 1938. A comparison of these two totals shows that there was an increase of \$.71 in the combined prices of three sizes

of coal. The relative of aggregates index number shown in row (f) amounts to $104.0, \frac{18.52}{17.81} \times 100$, and indicates that the prices increased 4.0 per cent from the base period. The aggregative index number cannot be compared with the other two for it is expressed in dollars. The two relative indexes can be compared, however, for they are both ratios. This comparison indicates that the use of the relative of aggregates and the average of relatives methods produces different results, namely, 104.0 and 104.2 respectively.

The reason for this difference becomes apparent after a study of the data in the table. The greatest actual increase occurred in the price of stoker coal. This increase of 50 cents together with the 19-cent increase in lump and the 2-cent rise in egg, amounted to a total increase of 71 cents. It is this 71 cents that is 4.0 per cent of the base period aggregate of \$17.81, $\frac{71}{17.81} \times 100 = 4.0$. In any aggregative index all of the price changes, whether they are large or small, or whether they occur in a commodity which has a high or low base price, are aggregated and related to the base aggregate. No attention is paid to the relative size of each base period price. On the other hand, in the average of relatives each base period price is of great importance. It was the lowest-priced coal that had the 50-cent increase, and this caused the large relative, 109.3, which is mainly responsible for the average of relatives index being larger than the relative of aggregates. If the 50-cent increase had occurred in the lump size and the 19-cent increase had taken place in stoker coal, the average of relatives index would have been lower than 104.2; whereas the relative of aggregates index number would not have been affected, for the two aggregates would have remained unchanged.

The two methods should be used on different occasions: (1) The relative of aggregates method gives greater weight to the higher priced units; it should be used when the actual dollar differences are considered most important or when all constituent commodities are at essentially the same price level. (2) The average of relatives method should be used when the relative price changes of the separate items are important and when there is no great diversity among the several relatives; when the prices of the constituent items are markedly different in size; or when it is difficult to obtain prices for commodities in comparable units. The average of relatives method gives greater weight to changes in low prices than to the same absolute changes in higher prices.

Weighted Index Numbers

Up to this point the composite index numbers described have all dealt with price data. They have been combined without paying any conscious attention to differences in the importance of the individual items included in the index. Like the man who, during the war, claimed that he made his excellent sausages from equal parts of horse and rabbit—one horse and one rabbit, we have put various items together without regard to the relative significance of each. In practice, however, it cannot be assumed that the individual items are equally important. Whenever prices or other items are combined in an index number, the relative importance of each must be taken into account and weights must be assigned accordingly. If it is decided that *equal* weights should be given to each, then the methods of unweighted indexes may be applied, but if unequal weights are needed one of the weighted methods must be employed. In reality, therefore, no composite index is unweighted. If a set of weights is not consciously applied, each element of the index automatically has *some* system of weighting.

The three methods which have already been explained form the basis for the construction of weighted index numbers. Or, to be more exact, from the haphazard weighting methods, the purposely weighted index numbers may be developed. Each method of construction will be illustrated by the use of examples.

The Weighted Aggregative Index.—The weighted aggregative index number is obtained by multiplying the price of each commodity by a number representing the weight and adding the resulting products. The weights used in an aggregative price index number are usually quantities. As a result, the product of price times weight gives the value of a certain quantity of each commodity included, and the sum of the individual values is the weighted aggregative index number. This number represents the total amount of money necessary to buy the indicated quantities of the several commodities at the prices which were effective on a given date.

This method of construction is employed by the Bureau of Business and Social Research at the University of Buffalo in preparing its monthly index of retail food prices. Once each month, prices are collected for forty-two food items commonly used in Buffalo homes. Each price is weighted by a number that represents the amount of that commodity which is consumed by the average family in a year. These

42 products are then added and the resulting aggregate is the total cost or value at current prices of the quantities of the 42 commodities necessary for a family for one year. This total is the weighted aggregative index number for that month. When this aggregate is compared with the total cost of exactly the same amounts of the identical items for a different month, the direction and extent of the change in retail food prices is apparent.

TABLE 100
WEIGHTED AGGREGATIVE INDEX OF RETAIL FOOD PRICES
IN BUFFALO, AS OF THE TENTH OF EACH MONTH,
1937 AND 1939 *

MONTH	1937	1939
January	\$403.30	\$377.30
February	401.77	373.82
March	403.86	371.32
April	414.18	365.25
May	421.33	363.64
June	421.98	363.02
July	410.41	370.34
August	419.05	367.02
September	409.52	372.98
October	415.61	374.08
November	403.81	369.89
December	401.03	364.09

* Bureau of Business and Social Research, University of Buffalo.

For instance, in Table 100 the total of the values of the indicated quantities of the 42 commodities in Buffalo amounted to \$403.30 in January, 1937. The total value of the same quantities of identical commodities in January, 1939, two years later, was \$377.30. In this twenty-four-month period, the cost of a year's supply of the forty-two commodities decreased \$26.00. Which prices have gone down, however, or the amount of the decrease in any individual price, cannot be ascertained from this index number. In fact, prices of some of the 42 items may have increased, and some of them may have remained the same. But the net change, i.e., the total amount of the decreases less the total amount of the increases indicates a generally decreasing tendency in the prices of these items.

The Relative of Weighted Aggregates Index.—The relative of weighted aggregates index number is derived from the weighted aggregate, just as in the case of the corresponding unweighted indexes. The weighted aggregate for a certain year is selected as the base for the aggregates of all the other years. The resulting quotients are the relative of weighted aggregates index numbers.

Table 101 illustrates the effect of applying weights to the prices that were shown in Table 97. The weights represent the quantities of each of the three commodities consumed by an average family in a week: 10 one-pound loaves of bread, 2 pounds of butter, and 12 quarts of milk.

TABLE 101
COMPUTATION OF WEIGHTED AGGREGATIVE AND RELATIVE OF
WEIGHTED AGGREGATES INDEX NUMBERS OF PRICES OF THREE
FOOD COMMODITIES IN THE UNITED STATES, FOR THE
MIDDLE OF JANUARY, 1938 AND 1939

COMMODITY	UNIT	(1)	(2)	(3)	(4)	(5)
			JAN. 18, 1938		JAN. 17, 1939	
		WEIGHT (Number of Units Consumed in One Week)	AVERAGE PRICE PER UNIT (cents)	COST OF WEEK'S SUPPLY (1) × (2) (cents)	AVERAGE PRICE PER UNIT (cents)	COST OF WEEK'S SUPPLY (1) × (4) (cents)
a) Bread, white	lb.	10	8.9	89.0	8.1	81.0
b) Butter	lb.	2	40.4	80.8	33.5	67.0
c) Milk, fresh delivered..	qt.	12	12.7	152.4	12.7	152.4
d) Total = weighted aggregative index				322.2		300.4
e) Relative of weighted aggregates index, row (d), col. 3 and 5 ÷ col. 3, × 100...				100.0		93.2

In this case, the weighted aggregative index numbers, row (d), indicate that the specified quantities of each of the three commodities could be purchased for 21.8 cents less on January 17, 1939, than on January 18, 1938. The relative indexes, row (e), indicate that in January, 1939, the combined cost of the three items was 6.8 per cent less than a year earlier.

Another illustration of a similar calculation is given in Table 102 which compares the results of the monthly weighted aggregative and the relative of weighted aggregates index numbers of retail prices of food in Buffalo for the year 1939.

These two sets of index numbers can be interpreted in different ways. Column 1 shows that the general trend of retail prices during the year 1939 was downward from \$377.30 in January to \$364.09 in December. There was a range of \$14.28 between the low of \$363.02 in June and the high of \$377.30 in January.

From the indexes in column 3 it can be seen that between July and December prices declined 1.66 points on the January base value ($98.16 - 96.50 = 1.66$). It should be observed that this is a decrease of 1.69 per cent. ($1.66 \div 98.16 \times 100 = 1.69$).

TABLE 102

COMPUTATION OF THE RELATIVE OF WEIGHTED AGGREGATES FROM THE
WEIGHTED AGGREGATE INDEX NUMBER, RETAIL FOOD PRICES IN
BUFFALO AS OF TENTH OF EACH MONTH, 1939

MONTH 1939	(1) WEIGHTED AGGREGATIVE INDEX	(2) COMPUTATION OF RELATIVE OF WEIGHTED AGGREGATES *	(3) RELATIVE OF WEIGHTED AGGREGATES INDEX NUMBER (JAN 1939 = 100)
January	\$377.30	$\$(377.30 \div \$377.30) \times 100$	100 00
February	373 82	$(373.82 \div 377.30) \times 100$	99 08
March	371.32	$(371.32 \div 377.30) \times 100$	98.42
April	365 25	$(365.25 \div 377.30) \times 100$	96 81
May	363.64	$(363.64 \div 377.30) \times 100$	96.38
June	363.02	$(363.02 \div 377.30) \times 100$	96.22
July	370.34	$(370.34 \div 377.30) \times 100$	98.16
August	367.02	$(367.02 \div 377.30) \times 100$	97.27
September	372.98	$(372.98 \div 377.30) \times 100$	98.85
October	374.08	$(374.08 \div 377.30) \times 100$	99.15
November	369.89	$(369.89 \div 377.30) \times 100$	98.04
December	364.09	$(364.09 \div 377.30) \times 100$	96.50

* As suggested on page 466, this computation may be shortened by finding the reciprocal of 377.30 and multiplying

The Weighted Average of Relatives Index.—In computing a weighted average of relatives, the weights are applied to the relatives of the prices and not to the prices themselves. The relatives are abstract numbers; therefore they must be multiplied by weights expressed in a common unit or the resulting products cannot be totaled. Dollar value is the only common unit of such diverse commodities as pounds of beef, bushels of wheat, quarts of milk, etc. Therefore value weights must be used in computing an average of relatives index.

In Table 103 both the unweighted and the weighted average of relatives indexes are computed for the same price data that were used in Table 101. Rows (e) and (f) show the indexes: columns 3 and 7 are the unweighted averages of relatives constructed by the same process illustrated for coal prices in Table 99; columns 4 and 8 are the weighted average of relatives indexes for the base period, January, 1938, and for January, 1939. The weighted indexes are obtained by dividing the sums of the weighted relatives, 322.0 and 300.1, by the sum of the weights, \$3.22. The weighted index number in January, 1939, is 1.9 index points higher than the unweighted number.

TABLE 103
COMPUTATION OF AVERAGE OF RELATIVES INDEXES, WEIGHTED AND UNWEIGHTED, OF PRICES OF THREE FOOD COMMODITIES
IN THE UNITED STATES, FOR THE MIDDLE OF JANUARY, 1938 AND 1939

COMMODITY	UNIT	(1) NO. OF UNITS CONSUMED PER WEEK	(2) JANUARY 18, 1938			(5) VALUE WEIGHTS AS OF 1/18/38 (1) × (2)	(7) JANUARY 17, 1939		(8)
			PRICE	RELATIVES (2) ÷ (1) × 100	WEIGHTED RELATIVES (3) × (5)		PRICE	RELATIVES (6) ÷ (2) × 100	
a) Bread, white	lb.	10	\$.089	100.0	39.0	\$.89	\$.081	91.0	81.0
b) Butter	lb.	2	.404	100.0	81.0	.81	.335	82.9	67.1
c) Milk, fresh, delivered	qt.	12	.127	100.0	152.0	1.52	.127	100.0	152.0
d) Total	...	...	...	300.0	322.0	\$3.22	...	273.9	300.1
e) Unweighted index, row (d), cols. 3 and 7 ÷ 3				100.0	...	...	...	91.3	...
f) Weighted index, row (d), cols. 4 and 8 ÷ col. 5				...	100.0	...	...	...	93.2

Formulas for the Basic Methods

The six methods of computing index numbers (three unweighted and three weighted) can be expressed in symbols. The system of symbols and formulas which follows is set up for the computation of price indexes. With proper interchange of symbols the formulas can be used for quantity indexes.

SYMBOLS

p = the price of an individual commodity.

q = the quantity of an individual commodity.

Subscripts

k = the period for which the given index is being computed, year, month, etc.

o = the period (month, year, or average of several years) used as the base of the index.

r = the period of the prices used in weighting an index.

s = the period of the quantities used in weighting an index.

Superscripts

$\prime \prime \prime \dots \dots \dots ^{(n)}$ indicate the first, second, third and so on to the n th commodity included in an index.

p_o = any individual price in the base period.

p_k = any individual price in any k th period.

p_r = any individual price in any r th period used in the weights.

q_o = any individual quantity in the base period.

q_k = any individual quantity in any k th period.

q_s = any individual quantity in any s th period used in the weights.

UNWEIGHTED FORMULAS

1. Aggregate

$$p'_k + p''_k + p'''_k + \dots + p_k^{(n)} = \Sigma p_k$$

2. Relative of Aggregates

$$\frac{p'_k + p''_k + \dots + p_k^{(n)}}{p'_o + p''_o + \dots + p_o^{(n)}} = \frac{\Sigma p_k}{\Sigma p_o}$$

3. Average of Relatives

$$\frac{\frac{p'_k}{p'_o} + \frac{p''_k}{p''_o} + \dots + \frac{p_k^{(n)}}{p_o^{(n)}}}{n} = \frac{\Sigma \frac{p_k}{p_o}}{n}$$

WEIGHTED FORMULAS

4. Weighted Aggregative

$$p'_k q'_s + p''_k q''_s + \dots + p^{(n)}_k q^{(n)}_s = \Sigma(p_k q_s)$$

5. Relative of Weighted Aggregates

$$\frac{p'_k q'_s + p''_k q''_s + \dots + p^{(n)}_k q^{(n)}_s}{p'_o q'_s + p''_o q''_s + \dots + p^{(n)}_o q^{(n)}_s} = \frac{\Sigma(p_k q_s)}{\Sigma(p_o q_s)}$$

6. Weighted Average of Relatives

$$\frac{\frac{p'_k}{p'_o}(p'_r q'_s) + \frac{p''_k}{p''_o}(p''_r q''_s) + \dots + \frac{p^{(n)}_k}{p^{(n)}_o}(p^{(n)}_r q^{(n)}_s)}{p'_r q'_s + p''_r q''_s + \dots + p^{(n)}_r q^{(n)}_s} = \frac{\Sigma\left[\frac{p_k}{p_o}(p_r q_s)\right]}{\Sigma(p_r q_s)}$$

or

$$\frac{p'_k}{p'_o} W' + \frac{p''_k}{p''_o} W'' + \dots + \frac{p^{(n)}_k}{p^{(n)}_o} W^{(n)} = \Sigma\left(\frac{p_k}{p_o} W\right)$$

in which

$$W' = \frac{p'_r q'_s}{\Sigma(p_r q_s)}; \quad W'' = \frac{p''_r q''_s}{\Sigma(p_r q_s)} \quad \text{etc.}$$

The subscripts r and s have been suffixed to the weights in formulas 4, 5, and 6 to make provision for the use of quantities or values for whatever period is deemed suitable in a particular construction. Specifically in formula 6 the notation is flexible enough to include the case in which the prices and quantities are for different periods ($r \neq s$) as well as the case in which values of a certain period are used ($r = s$). If weights for the current period are used, $r = s = k$; if base period weights are used, $r = s = o$. In the latter case or any other in which $r = o$, formulas 5 and 6 give identical results and the question of which to use will depend upon whether the weights are more readily available as values or as quantities.

The alternate form of formula 6 differs from the first form in that the weights are changed from dollar amounts to a distribution totaling unity, or they can be expressed as per cents totaling 100. The price relatives are then multiplied severally by these fractional weights. The two methods of computation are, of course, equivalent.

PROBLEMS OF INDEX NUMBER CONSTRUCTION

The selection of the most appropriate method is an important step in the computation of an index number, but there are in addition

a number of other problems which must be considered. The average user of index numbers needs to be aware of these problems so that he may in some measure be able to appraise critically the methods employed in the construction of the indexes which he uses, and on that basis to judge their validity for his particular purpose. These major problems include: (1) the purpose; (2) the items to be included; (3) the choice of the base period; (4) the selection of weights; and (5) the kind of average to be used.

The Purpose

The purpose an index number is intended to serve and the reasons for its development determine to a large extent the method of construction that will be used. The kind of items to be included as well as their number, the base period, the weighting system, and the averaging process likewise depend upon the purpose. The user of an index number must therefore have some knowledge of its original purpose, and the producer of an index number must know and should be able to state specifically what his purpose is before he starts his calculations.

For instance, the Index Number of Employment regularly computed by the Bureau of Business Research of The Ohio State University has been developed to indicate the relative monthly changes in the number of persons employed in manufacturing industries in the state of Ohio. Since it is based on the number of persons on the payrolls of manufacturing concerns on a corresponding date each month, it is an *employment* index only and cannot be used by itself as a measure either of unemployment, payrolls, or man-hours worked. It is not possible to use any such index numbers of employment as indicators of unemployment, for they are constructed so as to measure only the relative changes in the number of persons actually employed. They do not take account of possible shifts in the total population of working age, and hence give no indication of relative changes in the number of persons not working. In similar fashion, the Index Number of Retail Food Prices of the United States Bureau of Labor Statistics was developed to indicate relative changes in food prices at retail. It should not be used to measure prices in general, or any level of business other than retail trade.

The published names of index numbers usually offer the best means of narrowing the search for a particular series, but before any index number is adopted to serve a certain need further investigation should

be conducted relative to its original purpose. The formal name of an index number does not always reveal this purpose. Consequently to use an index number simply because its name sounds pertinent may lead to misrepresentation or misinterpretation.

If a single index number proves inadequate, the use of several related indexes may fulfill a given requirement. For instance, if it is desired to obtain a summary of the course of business during the past year in a certain area, it may be difficult to find a single index that will suffice, unless "business" is narrowly and specifically defined. The general situation may be revealed, however, by a number of simple index numbers of series such as employment, payrolls, bank debits, construction contracts awarded, and car loadings. These should be compared with a composite index number of business activity if there is one available for that area.

Items To Be Included

The selection of the items to include in an index number presents two major questions for consideration: (1) how many different items should be included; and (2) which items should be included. Although these can be recognized as separate problems, they are so closely related that they will be treated simultaneously. In determining both the number and the specific items to include, there are three guides: (1) the purpose; (2) the location and accessibility of the data; (3) the tenets of accepted statistical practice.

The Purpose as a Guide.—A comprehensive and detailed statement of the purpose serves as a guide to the kinds of items that should go into an index. This statement will indicate certain limitations, much the same as those involved in "defining the problem"⁸ as a first step in planning an investigation. It is not enough, for example, to decide that a retail food price index is to be prepared. That statement provides three points of definition: it is to be a price index; the prices are to be at the retail level; and only food items are to be included. However, it is also necessary to know (1) the area to be included, (2) the economic class of people to be represented, and (3) the time period to be covered.

Food habits vary in different geographical areas. The food items commonly included in the workingman's diet in Europe are very different from those in this country. Less striking yet marked varia-

⁸ Chapter IV, p. 53.

tions in diets appear in the several sections of the United States. The constituents of a retail food price index number, therefore, should vary in accordance with the geographical section being represented; or, if the same items are included in all sections, their weights should be changed to correspond to the differences in consuming habits.

The economic class of people which the index number is to represent is an important criterion in selecting the food items to be included in the index. Foods are of many kinds and it would be possible to include both common and rare foods, such as cabbages and avocados, eggs and caviar. But an index intended to show changes in prices of food purchased by working class families should contain only those items commonly used by that group. To determine what the common items are in various income groups is one of the main reasons for the periodic studies of consumption habits conducted by the United States Bureau of Labor Statistics.

The time period covered is the final consideration because it is likely to depend on some of the other factors. In the United States during the last twenty years there has been a great increase in the number of food products which have become common in the diets of working families. It would probably be more feasible, therefore, not to go back too many years in preparing a food price index, as it would be difficult to select a set of items that would be equally representative during the entire period. The recognition of these changes is one of the reasons which led the United States Bureau of Labor Statistics in January, 1921, to increase the number of items in its index of retail food prices from 21 commodities to 42, and gradually to 84 in 1935.

The Location and Accessibility of Data.—The data for an index number may be available in necessary form, but may have to be collected from widely scattered sources. When this is the situation, the only limitation is usually the cost of collection. These costs must include the maintenance of a satisfactory recording system for accumulating the data, and for retaining their confidential nature if they have been collected on the promise of secrecy. The number of items or the amount of data included may be dependent entirely in this case upon the relation between the cost of collection and the funds available for the purpose.

In many cases data are already available for use but they have been collected casually or for other purposes. That is, data may have been

collected as an incident to some purpose which is foreign to later needs. Data so collected may not be entirely satisfactory for index number construction, but since "beggars can't be choosers" the desired index number must sometimes be modified in order to utilize data which can be obtained quickly and economically.

The general form of such data may appear to be satisfactory but they may prove unusable because of failure in the collection to adhere to rigid definitions, specifications, and standards, or because of unsatisfactory counting units. These deficiencies are usually of vital significance, and in any case they must be known before making a decision to disregard them. For instance, if wheat prices are desired, it is important to know whether they are prices for the same kind and grade of wheat in different markets at each point in time. As this is being written, for example, winter wheat, No. 2 hard, is selling for \$.73 at Kansas City and \$.87 in the New York market; spring wheat, No. 1 Dark Northern, is selling for \$.76 at Minneapolis, while No. 3 grade winter wheat in Chicago is selling for \$.72. The same kinds of differences are found in almost every field. In Columbus, Ohio, for instance, No. 1 Maine potatoes are \$.27 a peck, while No. 1 Idaho cobbles are \$.35. Although they are both white potatoes, these two varieties have different characteristics which explain the price differences. Lack of standards in naming and grading commodities, either from one place to another, or from time to time, may have an important bearing on the inclusion of available prices in an index number.

A further consideration is the speed with which certain data become available. If an index is being prepared to interpret current business conditions as promptly as possible from month to month, or from week to week, only those items that are obtainable on the date required can be included. This necessity frequently causes the elimination of items that would otherwise be chosen, but there is no alternative except to rely on the most satisfactory possible substitute.

Tenets of Statistical Practice as a Guide.—The third guide in choosing the items to include in an index arises in the statistical requirement that the prices or other data must provide a representative sample. The principles for selecting a sample have been treated in detail in chapter V, and at this point it is necessary to emphasize that the data collected for constructing index numbers must conform to these principles. The items must be of such a variety that they satisfactorily represent the known characteristics of the field being studied. The

number of items needed will depend on the degree of selection possible in choosing the sample, i.e., a large number will be needed for a random sample, whereas a relatively small number of items that are known to be representative should produce a reliable index number.

The Base Period

In the examples used in illustrating the various methods of construction, the first period of the index was always used as the base, but this is by no means standard practice. The base of an index may be any single period or combination of periods that provides the most suitable standard for comparison.

Choice of the Period.—There are a number of criteria for the selection of the base period for index numbers. The most important of these may be listed as, (1) normality of the period; (2) the ease of reliable recall of the conditions of the period; (3) trustworthiness of the data in the period; and (4) comparability of the period with that used in other index numbers.

It is frequently held that the base period should be a period that is "normal" or "average" with respect to all of the data of the series. There is no clearly defined description of normal or average as used for this purpose, but such terms usually mean a period when the average of the data appears at a level about midway between the recent highest and lowest values. A period of very high prices, for instance, should not be used as the base because the resulting index numbers of all other periods would be low by comparison. In contrast, if a period of very low prices is used as the base, the index numbers of prices in all other periods appear high in the series. Thus, neither a very high year, as 1929, nor a very low period such as that of the depression trough, 1931, 1932, and 1933, are suitable base periods, but the single year 1926, the three-year average 1923–25, or the five-year average 1935–39, are often considered normal periods. For the construction of an index number which includes many series of data, there is never a single year when every series included is normal in this respect—every year is likely to be abnormal for some series. There are bound to be abnormalities in the data, even though a period of several years is chosen for the base. The period in which abnormal situations are at a minimum should therefore be selected as the base for the index number.

There is danger that readers will interpret "normal" as meaning

the level at which the index *should* stand, instead of where it actually does stand at any given period. Every effort should be made in presenting index numbers to eliminate this misconception and to make clear that the acceptance of a "normal" criterion for the base period implies no moral obligation or sense of approval. It is, therefore, advisable to call the period chosen as the standard merely the base period, or 100 per cent, and not even to refer to it as "normal," because of the misleading popular connotation of the word.

A second criterion of the choice of a period for the base of an index is the ease with which the users of the number will be able to recollect the conditions which existed during the period. There is a tendency among many people to think of past periods as "the good old days," which is detrimental to the objective recollection of conditions which actually existed. If a period within easy memory is adopted as the base for an index number it has the psychological effect of seeming to give validity to the results. The use of this criterion demands that the base period be changed at least once in a generation.

Trustworthiness of data is another important criterion in establishing a base period. Greater attention to the accurate and comprehensive collection of data has resulted in successively more trustworthy compilations, so that a recent period is more likely to provide a reliable base than an earlier period. The United States Bureau of Labor Statistics Wholesale Price Index furnishes an example of a base period established by this criterion. Between 1902 and 1913 wholesale price changes were measured in terms of the average price for the ten years 1890 to 1899. In 1915, however, the base was shifted to the latest available year (1914) in order to utilize "the latest and most trustworthy price quotations as the base from which price fluctuations were to be measured, and second, to permit the addition of new articles to those formerly included in the index number."⁹ The policy of using the latest completed year was continued through 1914 and 1915. When the index number for the period 1917-19 was being prepared, however, it became clear that due to the instability of the war period a pre-war base should be chosen. Accordingly 1913, the last complete year before the World War, was chosen as a fixed year base. Greater comparability was thus also made possible because other index number series prepared by the Bureau of Labor Statistics were on a 1913 base. When the Wholesale Price Index was revised in 1927, the base was

⁹ Mitchell, *op. cit.*, p. 117.

changed to 1926. "This choice was made because of the fact that 1926 was the last completed year when the work of revising its series of wholesale price index numbers was undertaken by the bureau in the summer of 1927, and it therefore furnished the most dependable standard for measuring price changes. Moreover, taken as a whole, market conditions in 1926 were regarded as fairly close to normal for the post-war period."¹⁰

In the absence of other positive criteria, the base for a new index number is sometimes chosen solely because it is used by existing index numbers with which the new one is most likely to be compared. Ordinarily the base should not be selected in this way, for it may lead to the masking of important peculiarities of the data in the new index. However, it should be noted that, whatever the variety of reasons for differences in base period, index numbers are not directly comparable unless their base periods are identical. For this reason it is particularly important in the publication of index numbers that the base period always be indicated.

The Length of the Base Period.—Periods as short as one month and as long as ten years can be found in published index numbers. The Fairchild Index of Prices of Department Store Goods is based on prices in December, 1930. As mentioned above, the United States Bureau of Labor statistics at one time used 1890–99, as a base for its Wholesale Price Index. The bureau uses 1923–25 as a base for its Employment and Payroll Indexes and 1935–39 for its Index of Retail Food Prices. Other base periods commonly employed can be found by examining the *Survey of Current Business*, the *Federal Reserve Bulletin*, or *Standard Trade and Securities Statistical Bulletin*.

There is no fixed rule for the length of a base period. The problem of length of period can only be solved by considerations of the purpose of the index and the accessibility and characteristics of the data. A shorter period base is more likely to be affected by abnormal variations which may occur in the data and a longer period base is likely to be a better representation of the "normal."

The following short description of the considerations of a Swedish bank in changing the base period of its price index is illustrative of base period problems in general.

¹⁰ *Wholesale Prices, 1913 to 1928* (Bulletin No. 493 of the Bureau of Labor Statistics, United States Department of Labor, Washington, D. C., August, 1929), p. 2.

NEW BASIS OF CALCULATION FOR THE SVENSKA HANDELSBANKEN'S
PRICE INDEX NUMBERS

In the present issue the Bank initiates a new basis for calculating its index numbers. The pre-war period has obviously become less and less suitable as a standard of comparison, and already in 1928 the basic series was supplemented in the Index with two parallel series based, the one on averages for the years 1923–25, when the price level was fairly constant, and the other on the prices in 1926, which year the U. S. Bureau of Labor Statistics introduced as basis, but which, through the effect the coal strike in England had on prices, was less suitable for purposes of comparison from the European point of view.*

After the depression in the beginning of the 1930's, the conditions cannot be said to have become relatively normal until 1935, in which year prices were in a fairly good state of equilibrium. In retrospect that year seems to represent a mean position between the trough of the depression in 1932 and the crest of the subsequent revival in 1937. For statistical purposes, therefore, the year 1935 has been adopted by, among others, the Swedish Board of Trade, and also elsewhere in the Northern Countries, as a gauge well suited to the present situation, and the Bank has decided to introduce the same basis for its index calculations. Such a change has seemed all the more advisable as certain qualities of goods included in the original index series, but which had gradually disappeared from the market, have had to be provisionally replaced by others. Moreover, a number of commodities, having acquired quite a different significance in the trade turnover compared with the period before the War, ought to be included or be given a revised weight in the calculations. Further, a careful inquiry has been made into the reliability of the material from which the price statistics are compiled, particularly in regard to the import and export figures.¹¹

* A good many index calculations, including those of the League of Nations, are nowadays based on the 1929 figures, which, however, have this drawback, that to a great extent they represent the maximum figures attained under boom conditions. Since the suspension of the gold standard, in order to be able to draw comparisons with the period immediately prior to the devaluation of the currency, calculations have also been made on the basis, as far as Sweden is concerned, of either September 1931 or the period October 1930–September 1931.

The Weights

Earlier in this chapter weights were defined and introduced in the index number calculation process. Here the problems of selection of weights, type of weights, shifting weights, and weight bias are discussed.

Selection of Weights.—Weights may be selected in such a way that they represent either the importance of specific commodities or the importance of certain kinds of fluctuation which are characteristic of

¹¹ *Index Svenska Handelsbanken Stockholm, Sweden, No. 153* (March 1939), pp. 11–12.

commodities in the same economic group. Mitchell points out that among raw materials, for instance, the prices of farm products as a group behave differently from those of mineral products.¹² Therefore, in developing an index of prices of raw materials, one might include in the farm products section the prices of a few of the more important commodities weighted to be representative of the entire group, rather than to include a large number of farm products and weight each one according to its own specific importance. This weighting process is somewhat analogous to the method of obtaining an arithmetic average of a sample distribution by using weights to represent the universe, as described in chapter XVI, pages 394–95. An example of this use of weights is found in the Index of Industrial Production computed by the Board of Governors of the Federal Reserve System which is described in the next chapter. The same procedure is found in the index numbers of prices 1913 to 1918 developed by the War Industries Board.¹³

The importance of the various items in an index number is determined by the purpose which the index is to serve. An index of the cost of living of wage earners' families, for instance, is constructed in order to show whether the expenses of the average wage earner's family are increasing or decreasing. Consequently the weights employed in such an index are usually per cents which measure the different major kinds of living costs in proportion to the total expenditures of the average family.

Physical Quantities or Values as Weights.—The factors used as weights for a given index number depend entirely upon the method of construction and the kinds of data being employed. If it is an index number of prices and one of the aggregative methods is used, i.e., a method which adds the actual unit prices, the measure of importance for weighting should always be *quantity* data of some kind, never value. Value is, in a sense, also a measure of price, since it equals price times quantity. Its use as a weight would actually have the effect of squaring the prices, which would give unwarranted importance to the larger price changes. For the same reason, an aggregative quantity index would be weighted by *prices*. For an average of relatives, on the other hand, *value* figures should be used as weights for the reasons explained on page 475. They provide the

¹² Mitchell, *op. cit.*, p. 66.

¹³ Mitchell, *ibid.*, p. 67.

best basis for measuring the respective importance of the several relatives that represent the changes of each separate component of the index number.

Whether the weights used will be quantities or values may, however, depend upon other considerations. For most kinds of commodities, quantitative data either of production or consumption have not been in existence over a very long period. Exchange values, on the other hand, have long been recorded in monetary units. In such cases two alternatives are possible: (1) use an average of relatives construction; (2) adjust values so that they approximate the missing quantity weights.

Other problems that are involved in determining how to measure the importance of components of an index number were recognized by Hudson in the following statement.

As in the case of most index numbers when the component series are not of equal importance, it becomes necessary to assign weights to the individual items in order that they may exert their due influence on the average. The determination of the relative importance of each item should not be left entirely to the discretion of the individual making the index but should rest, as far as possible, upon some objective standard. In the making of physical production indexes three standards of weighting have been used: The number of workers employed in the industry, the amount of power used, or the market value of the products ("value added" in the case of manufactures). The first two of these standards have largely been abandoned in favor of the third, which is no doubt a reasonable decision because market value is the most closely associated with production in the theoretical sense of utility creation.¹⁴

Constant or Variable Weights.—Index numbers are usually designed to show changes in the variable being measured—a price index, for instance, is usually designed to isolate changes in price from changes which may be due to other factors. None of the factors in the computation, except prices, should be allowed to fluctuate. The weights, therefore, should usually be kept constant for the period covered by the index. If prices and weights were allowed to vary simultaneously, the resulting index numbers would reflect changes due to both factors, and no one could tell what part of the final result was due to variations in price and what part was due to variations in the weights.

This raises the question, if the weights are to be held constant from

¹⁴ Philip G. Hudson, "The Technical Problems and Limitations to the Construction of Indexes of Physical Production," *Journal of the American Statistical Association*, Vol. 34 No. 206 (June 1939), p. 245.

the beginning to the end of the series, which specific period should they represent? In the examples used as illustrations of method, the weights were quantities or values as of the period used as the base of the index numbers, but this usage is not necessarily the most desirable one to follow in every case. It is somewhat easier to use base period weights than weights from some other period. Availability of data is frequently the determining factor in this decision and may require that base year weights be employed.

However, the importance of commodities may change during relatively short periods of time so that, if weights of an early period are used, there is a danger that the current index number will not accurately reflect the present relative importance of its several constituents. For instance, kerosene was an important item in the household budget twenty years ago but it might be excluded entirely from a modern index of cost of living for urban wage workers. Likewise, the cost of maintaining and operating an automobile is an important element in present-day cost of living although it did not exist twenty years ago for the rank and file of wage earners. If data are available, therefore, weights should ordinarily be chosen from a period recent enough to be representative of current conditions. When it is definitely known that basic changes in the importance of the constituents of the index are occurring, and if current and regularly reported data are available to measure these changes, revisions in weights may be introduced at regular intervals. Too frequent revisions, however, tend to impair the usefulness of an index number, so that ordinarily no change should be made as long as the weights continue to measure the approximate importance of the several components of the index. Examination of the practice of long-established indexes would indicate that weights have been changed at intervals of about eight or ten years.

Bias Due to Weighting.—Bias due to methods of weighting is almost certain to occur in some degree. In this sense "bias" means that, because of the failure of the weights to represent accurately the relative importance of the shifts in the elements included, the index number tends to understate or overstate the degree of change. Fisher has shown, for instance, that when applied in the weighted average of relatives method there is a downward bias imparted by the use of base-year value weights, and an upward bias by the use of given-year value weights.¹⁵ Bias may also result from a method of construction

¹⁵ Irving Fisher, *The Making of Index Numbers* (3d ed., Boston: Houghton Mifflin Co., 1927), pp. 96–108.

not suitable for the data, or from the use of an inappropriate kind of average.

The Type of Average Used

Any of the measures of central tendency discussed in chapters XVI and XVII may be employed in the average of relatives type of index number construction. There is no averaging in the aggregates. Significant differences in results may sometimes be completely concealed through failure to select the proper measure. The question of which measure to use is likewise related to the particular method of construction, i.e., the weighting system employed and the form of the data from which the relatives were computed. The relation of the characteristics of the several measures to the problems of index number construction and the advantages and disadvantages of each measure will be explained in some detail.

The Arithmetic Average.—The arithmetic average has a number of advantages in index number construction, the chief of which is ease of computation (see chapters XVI and XVII). Secondly, its method of calculation is familiar to most persons. It must be recalled, however, that this characteristic may become a drawback in that familiarity may breed contempt and indicate "a dangerous lack of curiosity rather than a clear understanding of the figures."¹⁶ Thirdly, the arithmetic average can be used in algebraic manipulation.

However, when the arithmetic average is used in index numbers it displays the same undesirable features already noted in chapter XVII. A few inordinately high relatives in a group of data may distort the average in an upward direction. This is an especially important characteristic, for in index number construction it is not desirable that a single price change should have this kind of biasing effect on the resulting index. Furthermore, the use of the arithmetic average may give contradictory results concerning the direction of price changes. The computation of an unweighted average of relatives in Table 104 illustrates this kind of difficulty.

If period 1 is used as the base, the arithmetic average of the relatives for period 2 is 125, showing that prices have increased by 25 per cent from period 1 to period 2. However, if period 2 is used as 100, then the arithmetic average of relatives of prices in period 1 is 125,

¹⁶ Mitchell, *op. cit.*, p. 73.

showing that prices were 25 per cent higher in period 1 than in period 2. The contradiction is due to the peculiar characteristic of the arithmetic average called by Fisher an "inherent tendency to exaggeration."¹⁷ This tendency toward an upward bias always exists when the arithmetic average is used in computing an index number, irrespective of the method of weighting.

TABLE 104
COMPUTATION OF INDEX OF HYPOTHETICAL PRICES USING ARITHMETIC AVERAGE

COMMODITY	PRICES		COMPUTATION OF ARITHMETIC AVERAGE OF RELATIVES INDEX			
			Period 1 as Base		Period 2 as Base	
	Period 1	Period 2	Per. 1	Per. 2	Per. 1	Per. 2
Milk (qt.)	\$.20	\$.10	100	50	200	100
Bread (lb.)	.05	.10	100	200	50	100
			2/ 200	2/ 250	2/ 250	2/ 200
Arithmetic average of relatives index number	..	..	100	125	125	100

The Median.—The median has distinct advantages in computing index numbers; nevertheless it has never become widely used for the purpose. It is the easiest of all the averages to compute when an index number includes a large number of items. Index numbers are samples purporting to represent a universe; hence sampling error is important. The median may have some advantage over the arithmetic average in this respect. In the conclusion of his study of the effects of different averages in the construction of index numbers, Fisher says: "The simple median, except when there are very few commodities, is probably at least as good on the average as a substitute for a weighted index number as is the simple geometric."¹⁸

The median has drawbacks, however, which may account for the infrequency of its use in index number construction. (1) In small samples, the median is likely to be erratic. When a small number of items is being used, there may be big gaps in the data which will seriously affect the value of the median. (2) The median cannot be manipulated algebraically, as can the arithmetic average, and conse-

¹⁷ Fisher, *op. cit.*, pp. 86 ff. Cf. pp. 29–30. "But we shall see that the simple arithmetic average produces one of the very worst of index numbers. And if this book has no other effect than to lead to the total abandonment of the simple arithmetic type of index number, it will have served a useful purpose."

¹⁸ Fisher, *op. cit.*, p. 264.

quently it cannot be used with ease in developing formulas. (3) Indexes computed from the median are not "perfectly reversible; that is, they cannot be shifted from one base period to another by simple division without ambiguity." (4) "If the number of commodities included in an index is even, the position of the median may be indeterminate, though within a determinate range."¹⁹

The Mode.—From the point of view of computation the mode possesses the same advantages as the median. Likewise its use removes the need for weights. Extreme relatives, either high or low, are totally disregarded in the mode, hence this source of bias is eliminated.

Far outweighing these advantages, and in fact removing the mode from further consideration, are the following disadvantages: (1) unless the index number includes a large number of items the distribution of the relatives may exhibit no clearly defined mode; (2) the mode is extremely insensitive to changes, hence an index number using it would measure only long time changes in conditions; (3) an index number with an early year base often exhibits a bi-modal tendency and as time progresses the individual relatives may become dispersed into numerous small subgroups with the result that no mode can be determined.

The Geometric Average.—The geometric average has merits which some statisticians credit most highly. In the first place, it reduces the danger of distortion in the index number produced by asymmetry in the distribution resulting from a few relatively large increases in the data. The characteristics of the geometric average make it peculiarly advantageous as a measure of central tendency of ratios, and hence it is especially valuable in constructing index numbers that involve the averaging of relatives.

Secondly, the advantage of the geometric mean in showing the true direction of change in a group of relatives is illustrated by applying it to the problem of Table 104, in which the arithmetic average failed. When the unweighted geometric average is used, it makes no difference which year is taken as the base, for, as shown in Table 105, there has been no change in the price index of the two commodities.

The disadvantages in using the geometric average, however, probably outweigh the advantages. First, this average is not generally known and consequently the results of its use are apt to be misinterpreted. Mitchell has pointed out that index numbers computed by using geometric averages do not have "any direct bearing upon changes

¹⁹ Mitchell, *op. cit.*, pp. 72-73.

TABLE 105
COMPUTATION OF INDEX OF HYPOTHETICAL PRICES USING GEOMETRIC AVERAGE

COMMODITY	PRICES		COMPUTATION OF GEOMETRIC AVERAGE OF RELATIVES INDEX			
			Period 1 as Base		Period 2 as Base	
	Period 1	Period 2	Per. 1	Per. 2	Per. 1	Per. 2
Milk (qt.)	\$.20	\$.10	100	50	200	100
Bread (lb.)	.05	.10	100	200	50	100
Geometric average ...	..	..	100*	100*	100*	100*

* These averages are computed by using the equation, Geometric Average =

$\sqrt[n]{p' \times p'' \times \dots \times p^{(n)}}$ which when applied becomes $\sqrt{100 \times 100} = 100$ and $\sqrt{50 \times 200} = 100$.

in the purchasing power of money," since they measure changes in the average ratios of changes in price rather than changes in the money costs of goods.²⁰ Second, geometric averages are much more difficult and time-consuming to compute than any of the other forms of averages. Third, the length of the computation makes it practically impossible to use this average in the preparation of an index number which must be available promptly.

TESTS OF INDEX NUMBERS

Time Reversal and Factor Reversal Tests

Fisher has suggested two tests, (1) the time reversal test, and (2) the factor reversal test, for determining the effects of bias in various methods of index number construction.

The time reversal test was devised to determine whether a certain method of construction would give corresponding results if the index number were calculated forward and backward between two given periods of time. For instance, to meet the time reversal test an index number that increases from 100 in 1938 to 150 in 1939 should show a decrease of $33\frac{1}{3}$ per cent if the same method is used in calculating 1938 from 1939 as a base. The simple geometric average of relatives is the only one of the average of relatives indexes that meets this test.

The factor reversal test can be applied only to weighted formulas. It may be stated simply as follows: for a single commodity, price $\times$ quantity = value; a composite index which possesses the same characteristic meets the factor reversal test. More specifically the test

²⁰ *Ibid.*, pp. 71-76

requires that when a price index and a quantity index of a given set of commodities are computed by the same formula for any given period on another period as base, the product of these two indexes shall be equal to the value index for the same periods. In symbols, using an aggregate formula and base period weights, the test is met, if

$$\frac{\Sigma(p_k q_o)}{\Sigma(p_o q_o)} \times \frac{\Sigma(q_k p_o)}{\Sigma(q_o p_o)} = \frac{\Sigma(p_k q_k)}{\Sigma(p_o q_o)}$$

None of the index numbers described thus far meets this test.

The "Ideal" Index Number

An index number formula which eliminates all bias and meets both the factor reversal and time reversal tests was developed simultaneously by several writers. Fisher's formula for this index, described as Number 353 in *The Making of Index Numbers*, and widely known as Fisher's Ideal Index Number,²¹ is:

$$\sqrt{\frac{\Sigma(p_k q_o)}{\Sigma(p_o q_o)} \times \frac{\Sigma(p_k q_k)}{\Sigma(p_o q_k)}}$$

The major disadvantage in this type of construction appears to be one of calculation. Few makers of index numbers can afford to employ this method because of the amount of time consumed in its computation.

Fisher's Ideal Index Number meets the two tests described, and it furnishes a satisfactory index of purchasing power. It does not follow that the use of the formula guarantees a reliable or even a good measure of price change alone. Simple methods competently employed with good data will frequently give more trustworthy results than highly refined methods.

Comparing Index Numbers—Base Shifting

When index numbers which have been constructed on different base periods are to be compared, it is necessary to "shift" one index to the same base period as the other, so that changes in the two will be measured from the same point in time. An index number can be shifted to a new base by dividing each of the index numbers in the series by the index number of the period selected as the new base. Although the method of shifting is simple, the results are apt to be misleading due to the introduction of bias.

²¹ For a complete description see Fisher, *op. cit.*, pp. 213 ff.

There is some variation between theory and practice on the question of base shifting. The aggregates can be shifted from one base to another without bias. Any index which meets the base reversal test can be shifted freely. The unweighted geometric is the only one of the average of relatives indexes that possesses this property, although the median usually produces little bias in the shifting process. But an arithmetic average of relatives may produce significant bias when the base is shifted. Nevertheless common practice affords precedent for such shifting when no alternative is available. In such cases the resulting comparison must not be considered as too precise because the shifted index is only approximately accurate.

Strictly speaking an index which is being shifted to a new base must be composed of the same items, including both quotations and weights, during the whole period of the index. Yet the most common use of base shifting is to link a current index containing one group of items to an earlier period parallel index containing a similar but not identical group of items. This procedure is legitimate if the old and new groups of items are equally representative of the same universe.

SPECIFIC USES OF INDEX NUMBERS AND THEIR INTERPRETATION

Deflating

Index numbers are used to deflate the effects of price changes. Everyone has heard the expressions "money wages," measuring wages in terms of the current value of money, and "real wages," meaning wages in terms of some constant comparable unit or in terms of the actual goods which can be purchased for a given sum of money. For instance, wages may have doubled between two given dates, so that a person who earned \$25 per week at the earlier date received \$50 per week at the second date. If the prices of all the commodities and services which he usually purchased also doubled, he obtained no advantage through the wage increase. The money wage increased, but the real wage did not change.

The adjustment for price change is made by dividing the series throughout by a price index, i.e., by dividing each item in the given series by the corresponding item of the price index. The process itself is a very simple one, but the major problem, the selection of the proper deflating index, is primarily a matter of judgment. The rule to be followed is: use an index number computed from the prices of the

commodities whose values are included in the series that is to be corrected. In some cases this rule can be followed very easily; in others complications arise. For example, department store sales should be corrected by the Fairchild Index of Prices of Department Store Merchandise. But this index is available only as far back as January, 1931, and no similar index is available for earlier years.²²

Suppose that a factory wished to know whether its wage scale was adequate to preserve the scale of living of its workers. The monthly payrolls divided by the number of full-time workers²³ would give a series of average full-time monthly wages over a given period. The monthly wage of a worker is the value of his purchasing power, that is, the amount available for maintenance. Therefore if the average wages were divided month by month by a cost of living index (price of maintenance), the resulting series would indicate whether the real wages were maintaining a uniform level of purchasing power. If the real wages were increasing, the employees would be in an advantageous position, but if they were declining the reverse would be true and labor difficulties could be expected.

It should be noted that a number of "purchasing power" indexes are published in the same sources that publish indexes of cost of living, farm prices, retail food prices, etc. These are simply reciprocals of the price indexes and may be used in their place for the purpose of deflating, by multiplying instead of dividing.

An example of the necessity for selecting different deflating indexes for different purposes is shown in Table 106. The money price of wheat received by farmers in the United States between 1920 and 1937 is deflated by two different index numbers, giving the "real" prices which would have obtained if the value of the dollar had remained constant. Column 4 represents the real price received for wheat if articles used in further farm production are purchased with the pro-

²² The author has found that a usable index for this purpose can be computed by taking the following three groups from the Bureau of Labor Statistics Index of Wholesale Prices and combining them with the weights as indicated:

Textile products	62
Hides and leather products	19
House furnishing goods	19
Total	100

²³ This statement depends upon the assumption that a large part of the laboring force works full time. Otherwise there would be no relation between the wages of a full-time worker and an individual family's income.

TABLE 106

DEFLATION OF PRICES OF WHEAT RECEIVED BY FARMERS IN THE UNITED STATES, 1920-37*

YEAR	(1) AVERAGE SEASON PRICE OF WHEAT PER BUSHFL RECEIVED BY PRODUCERS	(2) (3) INDEX NOS. OF PRICES PAID BY FARMERS (CALENDAR YEARS 1910-14 = 100)		(4) (5) DEFLATED PRICE OF WHEAT PER BUSHFL RECEIVED BY PRODUCERS DEC. 1	
		All Commodities Bought for Use in Production	Commodities Bought for Family Maintenance	For Production (1) ÷ (2)	For Family Maintenance (1) ÷ (3)
1910-14	\$.87	100	100	\$.87	\$.87
1920	1.83	174	222	1.05	.82
1921	1.03	141	161	.73	.64
1922	.97	139	156	.70	.62
1923	.93	141	160	.66	.58
1924	1.25	143	159	.87	.79
1925	1.44	147	164	.98	.88
1926	1.22	146	162	.84	.75
1927	1.19	145	159	.82	.75
1928	1.00	148	160	.68	.62
1929	1.04	147	158	.71	.66
1930	.67	140	148	.48	.45
1931	.39	122	126	.32	.31
1932	.38	107	108	.36	.35
1933	.74	108	109	.69	.68
1934	.85	125	122	.68	.70
1935	.83	126	124	.66	.67
1936	1.03	126	122	.82	.84
1937	.96	135	128	.71	.75

* *Agricultural Statistics* (1939) and earlier issues.

ceeds; column 5 represents the real price of wheat if articles used for farm family maintenance are purchased with the proceeds; both are expressed in terms of the average dollar in the period 1910-14. Thus in 1920 a bushel of wheat commanded \$1.05 worth of farm machinery as compared with \$.87 worth in 1910-14. But a similar bushel of wheat commanded only \$.82 worth of consumption goods as compared with \$.87 worth in 1910-14. There is much less variation in the "real" prices than in the money prices. And there is also less variation when the prices of wheat are deflated by an index of prices of goods used for farm family maintenance than when deflated by the index of prices of commodities used in farm production.

As measured in terms of goods used in further production, wheat prices have reached the 1910-14 purchasing power only in 1920, 1924 and 1925. As measured in terms of consumption goods, 1910-14 purchasing power has been attained only in 1925.

Interpretation

The interpretation of index numbers can be illustrated by the use of examples. For December, 1932, the United States Bureau of Labor Statistics reported an index of employment in manufacturing industries in the United States of 64.1. At the same time the index of payrolls in manufacturing industries was 46.1. Both of these indexes use the three-year average, 1923–25, as the base.

These figures indicate that as compared with the base period, the number of employees of manufacturing industries in the United States in December, 1932, had decreased by 35.9 per cent, and that payrolls were 53.9 per cent lower. The decline in payrolls was 50.1 per cent more than the decline in the number employed, i.e., $100(53.9 \div 35.9) - 100 = 50.1$.

A year later, in December, 1933, the index of employment was 69.0, and the index of payrolls was 48.5. From the average month in 1923–25 the number employed in manufacturing industries in the United States had declined 31.0 per cent while payrolls had declined 51.5 per cent. Conditions with respect to both the number employed and total payrolls improved from December, 1932, to December, 1933, however. Employment increased 4.9 index points ($69.0 - 64.1$) while payrolls increased by a smaller amount, 2.4 index points ($48.5 - 46.1$). This means that from December, 1932, to December, 1933, employment increased by 4.9 per cent of the average month in 1923–25, and in the same period payrolls increased by 2.4 per cent of the 1923–25 average month.

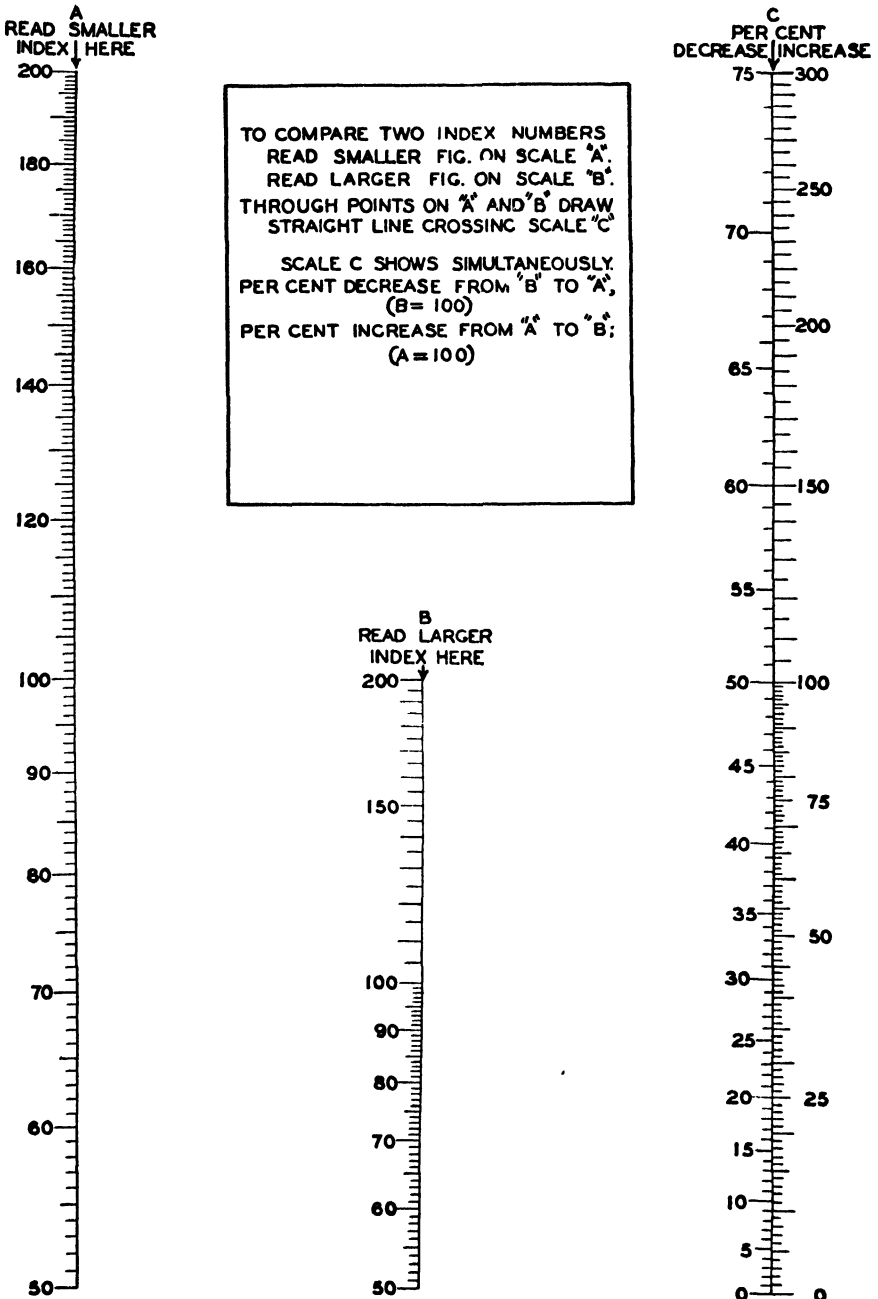
But in some ways, the latter statement is not a satisfactory comparison, for it does not indicate the percentage changes from December, 1932, to December, 1933. At this point there is very frequent misinterpretation. The difference in the number of points in the indexes at the two periods is sometimes erroneously considered as the percentage change, whereas, it is only a difference in index points.²⁴ In this case employment increased from 64.1 to 69.0, or 7.6 per cent, $100(69.0 \div 64.1) - 100$, between December, 1932, and December, 1933, and payrolls increased from 46.1 to 48.5 or 5.2 per cent, $100(48.5 \div 46.1) - 100$.

The indexes of employment and payrolls contain a sizable seasonal variation, being lowest in midsummer and highest in the fall and

²⁴ See chapter XI, p. 248 ff., for a review of this subject.

FIGURE 67

NOMOGRAPH FOR READING PER CENTS OF INCREASE OR DECREASE IN INDEX NUMBERS *



* A similar diagram appears in a special release of the National Industrial Conference Board published in June, 1933.

spring. This factor does not influence the particular comparisons discussed in the preceding paragraphs where the comparison is between the corresponding months of two different years. If on the other hand a comparison between June and December employment were wanted, the question of seasonal levels would have to be taken into account.

Since index numbers are time comparisons they actually contain several components. The preceding discussion has been concerned with the net result of *all* these components. Sometimes the seasonal component or the trend component or the cyclical component requires separate study. In other cases one of the components may be removed, as in the Federal Reserve Board's Indexes of (1) Industrial Production, (2) Department Stores Sales, (3) Employment. The methods employed in making seasonal and other adjustments are described in chapters XXI through XXVI. Chapter XXV contains a direct application of the methods of this chapter to data that have been analyzed by the methods of time series.

Use of the Nomograph

The nomograph of Figure 67 provides a simple means of obtaining percentages of increase or decrease, as may be illustrated by its use with the preceding example. Following the instruction on the graph, place a straight-edge on 69.0 on the middle line (*B*) and on 64.1 on the left line (*A*), and on the line (*C*) at the right read 7.6, approximately, which indicates that there has been a 7.6 per cent increase in the index from December, 1932, to December, 1933. The use of this nomograph will save a great deal of time in computation, and will avoid much confusion in determining the percentage change between any two points of an index series.

Conclusion

The index number is and will continue to be one of the most easily accessible statistical devices for business men. New applications or index numbers and developments in calculation methods as well as evaluation of known methods are being made continually. However, there are many possibilities of misinterpretation and biased results in their use. Careful study of methods and practice in handling them will increase the value of index numbers in business administration.

PROBLEMS

1. State the advantages of using index numbers rather than original data
2. Find in published sources an example, other than those listed in Figure 66, page 465, of **each of the four types of index number**. Give exact references
3. a) Using column 1 in the table below, compute, on 1929 base, an index of industrial employment in Buffalo.
b) Compare your index with that given in column 2 and discuss industrial employment during this period, in Buffalo as compared with the entire United States.

YEAR	(1) NUMBER OF EMPLOYEES ON PAYROLLS OF BUFFALO INDUSTRIES, 1929-39* (MONTHLY AVERAGES)	(2) INDEX OF INDUSTRIAL EMPLOYMENT IN THE UNITED STATES† (1932 = 100)
1929	59,349	161
1930	50,342	139
1931	41,217	118
1932	30,903	100
1933	32,016	111
1934	38,125	130
1935	41,015	138
1936	47,876	150
1937	52,818	165
1938	38,675	136
1939	43,461	147

* Data from Bureau of Business and Social Research, University of Buffalo.

† Derived from data in *Federal Reserve Bulletin* (January, 1940), and *Survey of Current Business Weekly* (February 1, 1940).

4. a) Compute indexes of prices by three methods from the following:

TYPES AND PRICES OF TOBACCO PRODUCED IN THE UNITED STATES

TYPE OF TOBACCO	SEASON AVERAGE FARM PRICE PER POUND			
	1926	1931	1935	1938
Flue-cured	\$.249	\$.084	\$.200	\$.222
Virginia fire-cured	.078	.047	.102	.107
Burley	.131	.087	.191	.190

- b) What are some of the objections to unweighted index numbers?

5. Compute an index number of retail prices from the following data:

ARTICLE	PRICE (per lb.)		NUMBER OF POUNDS CONSUMED PER YEAR BY AN AVERAGE FAMILY
	1913	1935	
Rib roast	\$.20	\$.24	31
Potatoes	.016	.02	704
Bread	.05	.08	531

6. Given prices and quantities as follows:

COMMODITY	PRICE (per lb.) 1937	QUANTITY 1937	PRICE (per lb.) 1941
Lead	\$.06	500,000 tons	\$.055
Tin	.60	80,000 long tons	.50
Copper	.12	800,000 tons	.14
Aluminum	.20	24,000,000 lbs.	.18

- a) Compute average of relatives price indexes for 1941 on 1937 as a base using (1) the weighting system, (p_{0q_0}), (2) the weighting system, (p_{kq_0}).
 - b) What sources of bias are likely to affect each index?
 - c) Compute a relative of weighted aggregates price index for 1941 on 1937 as a base.
 - d) What relation between the methods of construction of (a) and (c) is demonstrated by this problem?
7. Construct a weighted index of prices for December, 1934, on 1926 as a base using the following data. It is part of your task to determine from the data given what method of construction to use. Explain method used and why you selected this method.

Price of butter per lb. 1926—44 cents; December, 1934—31 cents
 Price of cheese per lb. 1926—23 cents; December, 1934—15 cents
 Price of evaporated milk per case, 1926—\$4.42; December, 1934—\$2.70
 Production of evaporated milk 1926—1,400,000,000 lbs.
 Exports of evaporated milk, 1926—76,000,000 lbs.
 Apparent consumption of cheese 1926—510,000,000 lbs.
 Average value of butter consumed 1925-27—\$75,000,000
 A case of evaporated milk weighs 45 lbs.

8. The following index was computed from the cost of materials used by the Q. Steel Company:

	(1) (2) PRICE PER TON		(3) (4) AMOUNT PURCHASED (thousand tons)		(5) (6) (1) × (3) (2) × (4)	
	1935	1940	1935	1940	THOUSAND DOLLARS	
Pig iron	\$18.00	\$22.50	3	5	54	112.5
Steel scrap	11.50	17.80	2	4	23	71.2
Bituminous coal	4.30	4.40	6	12	25.8	52.8
					102.8	236.5
Index					100.	230.

- a) What kind of index is this? Method?
 - b) What does it show regarding changes in prices of materials used? in quantities?
9. a) What kind of index is the following and what method of construction?
- b) Criticize the selection of items, weights, base period, and method.

INDEX OF AMOUNT OF TRAVEL BY PEOPLE OF THE UNITED STATES, 1930-40

	NUMBER (Monthly Averages)			RELATIVES		
	1930	1935	1940	1930	1935	1940
Revenue passengers carried one mile, steam railways (millions)	2,235	1,540	1,981	100	68.9	88.6
Passenger miles flown on scheduled airlines (thousands)	7,001	26,159	93,121	100	373.6	1,330.0
No. of visitors to national parks (thousands)	137	198	364	100	155.9	265.7
No. of passports issued	16,931	9,842	2,188	100	58.1	12.9
Total				400	656.5	1,697.2
Index				100	164.1	424.3

10. The following is an index of the consumption of fuels.

	(1)	(2)	(3)	(4)	(5)	(6)
	QUANTITIES (BbLS.)		PRICES (per bbl.)		(1) × (3)	(2) × (3)
	1946	1949	1946	1949	MILLION DOLLARS	
Fuel oil	24,000,000	26,000,000	\$.75	\$.85	18.	19.5
Gasoline	43,000,000	51,000,000	2.00	1.60	86.	102.
Kerosene	5,000,000	6,000,000	1.70	1.60	8.5	10.2
					112.5	131.7
Index					100	117

- a) What kind of index is it? What is the method used?
 - b) Name two different kinds of weighted indexes that can be computed from these data.
 - c) Compute the two indexes named in (b).
11. a) Enumerate and explain the principles of index number construction.
 - b) Only in exceptional cases will it be necessary for any of you to construct index numbers. Why, then, is it necessary to understand the principles of index number construction?
 12. Index numbers are samples, consequently care must be exercised to insure that the items included in the index are representative of the universe. Describe the universe represented by, (a) a wholesale price index, (b) a cost of living index, (c) an index of production of minerals, (d) an index of retail sales in rural areas.
 13. Discuss the advantages and disadvantages of using, (a) a single year or a period of years as the base of an index number, (b) a recent year or a remote year, (c) a depression period or a prosperity period.
 14. What is the relation of the "Ideal Formula" to the problem of constant or variable weighting?

15. In chapter XVI the geometric average was proposed as the proper measure of central tendency of ratios. Does it follow that the geometric average of relatives should be used in constructing index numbers? Discuss.
16. A student wishing to show the change in expenditures for advertising from 1929 to 1940 collected the following data:

"PRINTERS' INK" INDEXES OF ADVERTISING EXPENDITURES
(1928-32 = 100)

TYPE OF MEDIUM	INDEX OF EXPENDITURE	
	1929	1940
Magazines	125.4	82.4
Newspapers	118.9	79.0
Radio	71.5	347.4

He then proceeded as follows:

	1929	1940
Aggregate	315.8	508.8
Relative of aggregates index	100.0	161.1

- a) Do you agree that advertising expenditures had increased 61 per cent in 1940 as compared with 1929?
- b) Criticize anything that is unsatisfactory in the student's work.

REFERENCES

ARTHUR, H. B., "Weighted Aggregates and Index Numbers," *Journal of the American Statistical Association*, Vol. XXXII (June, 1937).

ASCHER, LEONARD, "Variations in Price Relative Distributions," *Journal of the American Statistical Association*, Vol. XXXII (June, 1937).

DAVENPORT, DONALD H., and SCOTT, FRANCES V., *An Index to Business Indices*. Chicago: Richard D. Irwin, Inc., 1937.

DAVIES, GEORGE R., "Index Numbers in Mathematical Economics," *Journal of the American Statistical Association*, Vol. XXVII (March, 1932, Supplement).

FERGER, WIRTH F., "Distinctive Concepts of Price and Purchasing Power Index Numbers," *Journal of the American Statistical Association*, Vol. XXXI (June, 1936).

FISHER, IRVING, "The Best Form of Index Number," *Quarterly Publications of the American Statistical Association*, Vol. XVII (March, 1921).

FISHER, IRVING, *The Making of Index Numbers*. Boston: Houghton Mifflin Co., 1927, third edition

- FISHER, IRVING, "The 'Total Value Criterion': A New Principle in Index Number Construction," *Journal of the American Statistical Association*, Vol. XXII (December, 1927).
- KING, WILLFORD I., *Index Numbers Elucidated*. New York: Longmans, Green and Co., 1930.
- LOVITT, W. V., "Index Number Bias," *Journal of the American Statistical Association*, Vol. XXIII (March, 1928).
- MITCHELL, WESLEY C., *Index Numbers of Wholesale Prices in the United States and Foreign Countries* (Bulletin No. 284 of the U. S. Bureau of Labor Statistics). Washington: Government Printing Office, 1921.
- MONTGOMERY, J. K., *The Mathematical Problem of the Price Index*. London P. S. King & Son, Ltd., 1937.

CHAPTER XX

SOME COMMONLY USED INDEXES

INTRODUCTION

IN THE preceding chapter, references to various indexes were introduced to illustrate particular points in the methods of index number construction. Explanations of the complete procedure involved in the development of several specific indexes have been reserved for this chapter.

The five indexes that will be described have been chosen because they illustrate the compromises that must always be made between the principles involved and the practical difficulties encountered in constructing any index number. The task of actually planning any but a simple index is one that practically never confronts a business man. However, he does often need to select from existing indexes the one best suited to his particular purpose. In order to choose discriminatingly and to understand the significance of the completed index he must be aware of the problems involved in each case and the methods applied in their solution.

These problems include not only the special techniques peculiar to constructing the index after the necessary data have become available, but also all of the factors that enter into the collection of representative sample data in the case of any statistical investigation. For this reason, the descriptions that follow are concerned with the preliminary choosing of sources of data and methods by which the original prices, quantities, etc., are combined into the main groups that appear in the index. The final application of a formula in the derivation of the combined index is in large measure dependent upon the earlier steps in the process.

The five indexes that will be described are: An Index of Cost of Living; An Index of Industrial Production; An Index of Factory Employment; An Index of Wholesale Prices; A Local Business Index.

AN INDEX OF COST OF LIVING

Every individual thinks of the cost of living as the total of the expenditures required to provide the necessities of life and the luxuries

which he feels he can afford at his level of income. Without actually keeping any accounting records, he usually knows, for example, that he is spending more for rent now than he paid last year; or that he seems to be living better than he did a year ago on the same total expenditure, that is, he has more or better food, or more comforts of life. He may have only a vague notion of all the factors involved, and yet his close contact with the situation provides some basis for his comparisons and the grounds for his conclusions and decisions.

The concept of the cost of living for a people or a nation must be more specifically defined, however. The cost of living is usually considered as the total of the expenditures for those goods and services which people consider absolutely essential to their maintenance during a given period of time. But there is no single level of income at which the amount of goods and services can be estimated. Before progress can be made, some standard or level must be accepted. The level is frequently called the "standard of living." It may be described in terms of the characteristics or manner of living of the group being measured; that is, by the income level of the group; the kinds of occupations engaged in by the "breadwinners" of the group, such as laborer, skilled, semi-skilled, "white collar," or professional; the size of family; or by combinations of these and other characteristics.

The actual costs of living for any one of these groups may be obtained in several ways. An early attempt to obtain costs of living, made similarly by LePlay in France and Engel in Sweden, involved living with artisan and peasant families in order to keep detailed records of the actual use of the families' incomes. At best, these records were very scanty and scarcely provided a basis for generalization concerning costs of living of a whole population.

A second method recently used in several government surveys is that of asking a large number of families having certain characteristics of income, age, size, etc., to keep detailed daily records of their actual expenditures over a specified period of time. These records are then averaged to obtain costs representative of the whole group for that period. This is an expensive method and can be used only at infrequent intervals of time; hence from this source continuous information concerning costs of living is not available.

A third method involves determining by a sample study the quantities of all the various kinds of commodities which families use during a specified period. These quantities of goods and services can then be

priced at any time to see what total dollar expenditure would be necessary to purchase them. This method permits a regular periodic and place-to-place determination of costs of living.¹ Even here the task is complicated by reason of the great variety of items included in the living requirements of families in the United States. Nevertheless, two major indexes of cost of living have been developed in the United States: one by the National Industrial Conference Board, published monthly, and one by the United States Bureau of Labor Statistics, published quarterly. The former will be described as an example of the methods involved.

National Industrial Conference Board

Index of Cost of Living in the United States, 1914-1939

Early in 1918, spurred on by the widespread demand for factual data concerning the changes in living costs resulting from the wartime price rise, the National Industrial Conference Board undertook the work of estimating the cost of living in the United States. While this preliminary compilation did not produce results that were intrinsically valuable, the experience gained more than justified the effort. Later in the same year, the Board inaugurated a systematic collection of data planned to provide an index which would appear at regular intervals.

Purpose

The Conference Board designed its index to "measure changes in the cost of living that occur from month to month."² The concept of cost of living as used by the Board was limited to expenditures of wage-earners in the United States, and these expenditures were divided into five major groups: food, housing, clothing, fuel and light, and sundries. Since wage-earners ordinarily live in cities, their habits of living and expenditures are peculiar to urban conditions and are different from those of agricultural or rural workers. Consequently, the index was constructed to represent urban conditions.

¹ This method was used in a study entitled, *Intercity Differences in Costs of Living*, by Margaret Loomis Stecker, Works Progress Administration, Research Monograph XII. Washington: Government Printing Office, 1937.

² M. Ada Beney, *Cost of Living in the United States, 1914-1936* (New York: National Industrial Conference Board, Inc., 1936), p. 13.

Kinds and Sources of Data

General.—The plan of the Conference Board Index depends upon monthly collection of the prices of most of the commodities consumed by urban workers' families. This requires an extensive price reporting service to secure proper representation of all sections of the country. The organization and operation of this collection service is described in some detail in the report previously cited. A summary of this phase of the work is presented in Table 107. There are 190 commodities listed and some of these are averages of individual prices. For example, in the subgroup, light, of the fuel and light group the price of gas is an average of the price of natural gas and manufactured gas. As indicated in the table, the prices in the several groups are reported from a number of cities. In some cases multiple reports are received from a single city. Thus as many as 25 stores in large cities report grocery prices.

TABLE 107
SUMMARY OF DATA AND SOURCES OF COST OF LIVING INDEX
NATIONAL INDUSTRIAL CONFERENCE BOARD

ITEM OF EXPENDITURE		No. OF COMMODI- TIES	No. OF REPORTING CITIES	REPORTING AGENCY
Group	Subgroup			
Food		84	51	Retail grocery stores
Housing		1	173	Real estate boards Chambers of Commerce Real estate agents Social agencies Individuals
Clothing	Men's	25}	93	Retail stores selling cloth- ing
	Women's	22}		
Fuel and light.	Fuel	3	95	Retail coal dealers
	Light	2	174	{ American Gas Association Edison Electric Institute
Sundries	Carfare	1	289	American Transit Associa- tion
	Drugs	12	14	A chain drug retailer
	Reading material...	2	94	{ Newspaper publishers American News Trade Journal
	Recreation	1	83	Motion picture theater operators
	House furnishings	30	93	Retail stores selling house furnishings
	Tobacco	3	..	A chain tobacco retailer
	Candy	4	14	A chain drug retailer

The last column of the table contains a list of the reporting agencies. The maintenance of regular relations with all of these co-operators, the replacement of those that withdraw for various reasons, and the establishment of new contacts for replacement and expansion is in itself a huge task.

The Conference Board does not collect food prices directly but uses the reports of the United States Bureau of Labor Statistics. Data for all of the other items are collected by the Board and all of the indexes are prepared by it. The magnitude of the work can be surmised from the following detailed description of the preparation of the housing index.

Housing Data.—The index of housing costs is described here because: (1) The problems are somewhat different from those faced in other indexes, and therefore provide further illustration of index number construction; and (2) since an index number of housing cost is not available for most localities, the methods described here may be followed for computing such an index number for any local area.

"Housing is represented in the Conference Board's index merely by rent."³ Anyone who has visited different parts of the United States will recognize that housing customs are not uniform. In some large cities, wage-earners commonly live in single detached houses; in other cities, they may live predominantly in two-family houses or apartments in multiple-family buildings. In some sections of the country, the cost of heat is customarily included in the rental price, and in other sections it is not. These differences, as well as numerous others, have caused several changes in the descriptions and definitions on the schedules for data collection, but at present, rentals are obtained on "four- and five-room houses, with bath, and four- and five-room flats, with bath, but without heat . . . (except in a few instances where heated flats are the prevailing type)."⁴ It is basically important, (1) that in each city, cost of rentals be collected for dwellings which are representative of the type customarily rented by wage-earners in that city, (2) that the rentals reported be for the same types of housing, continuously, and (3) that the data be collected in such a way that they will be representative of the relative changes in rentals paid by wage-earners.

Source: Each month, the Conference Board sends rent questionnaires to real estate boards, chambers of commerce, real estate agents.

³ *Ibid.*, p. 19.

⁴ *Ibid.*, p. 19.

social agencies, and qualified individuals. The questionnaire asks for the "average monthly rent" as of the fifteenth of the month, for the kind of housing described. When the form is sent out, the figure reported for the previous month is inserted so that the respondent can maintain comparability and avoid confusion.

Schedules are regularly sent to respondents in 173 different cities in the United States, ranging in population upward from 25,000, which taken altogether represent about one-third of the population of the United States. The number of schedules sent to each city is roughly proportional to the population of the city.

Method of combining data: The group index for housing is obtained as the end result of four separate computations: (1) the averaging of individual reports to give a city index, (2) the averaging of indexes of cities of similar size, (3) the averaging of weighted indexes for all cities, and (4) the shifting of the base from December, 1928, to 1923. These four steps will now be described in more detail.

The individual figure for the current month on each report for a given city is divided by the figure for the preceding month. These relatives expressed as per cents are averaged to give the relation between the two months of rents for the entire city. The rent index for the earlier month is then multiplied by the per cent relative to give the rent index for the current month. The index for the preceding month is on a December, 1928, base; hence by the last step the index for the current month is put on a December, 1928, base. Similar computation gives for each city a rent index on December, 1928, as a base.

The separate city indexes are combined in five group indexes according to size of city. For instance, to obtain the rental index for Group 1 in Table 108, containing cities of over 500,000 population, the indexes of 13 cities in the group are averaged; for group 2, the indexes of 22

TABLE 108

NUMBER OF CITIES IN EACH SIZE GROUP AND WEIGHTS
USED IN COMPUTING NATIONAL INDUSTRIAL CONFERENCE BOARD INDEX OF RENT

GROUP	POPULATION*	NUMBER OF CITIES	WEIGHT
1	500,000 and over	13	42.7
2	250,000 to 499,999	22	16.3
3	100,000 to 249,999	50	15.4
4	50,000 to 99,999	58	13.3
5	25,000 to 49,999	30	12.3

* Based on 1930 Census of Population.

cities are averaged, etc. A weighted average of the indexes of the city-size groups then is calculated to obtain the index of rents in the United States. The weights represent the proportion of the urban population living in cities of each size group, according to the 1930 census.

By this succession of averages, the rent index linked through percentage relatives to December, 1928, as the base period is finally obtained for the entire United States. To make this index useful as one of the five in the total cost of living, however, it must have 1923 as a base period. This shift is made by multiplying the current index on a December, 1928, base by the December, 1928, index on a 1923 base. The current month's rent index is then ready to be weighted and combined with the weighted indexes of the other four groups to make the Index of Cost of Living for the United States.

Base Period

When the Conference Board began the construction of an index it followed the lead of others by adopting July, 1914, as the base period. By 1930, however, the Board recognized that the pre-war base was an old and outdated standard of comparison. In its place the year 1923 was selected as a base because, "it represented the first post-war year of relatively stable economic conditions."⁵ Index numbers for all earlier periods were recomputed on the new base.

Weights

The weights used in computing the combined index are shown in Table 109, column 1. They were introduced in 1931 to replace a formerly used set that was no longer representative of the distribution of family expenditures. The new weights are based on a number of independent cost-of-living studies made in various parts of the country between 1921 and 1929. The percentage distribution represents the proportional expenditure for the five types of commodities. These per cents are therefore equivalent to value weights, but neither the prices nor the quantities are for any particular year.

The weighting systems used in preparing the group and subgroup indexes vary according to the form of the data and the auxiliary information available for weighting purposes. A complete presentation of this phase of the work appears in the original source. The calculation of the final index for June, 1936, is shown in Table 109.

⁵ *Ibid.*, p. 14.

TABLE 109
COMPUTATION OF COMBINED COST OF LIVING INDEX OF
NATIONAL INDUSTRIAL CONFERENCE BOARD, JUNE, 1936 *

EXPENDITURE GROUP	(1) WEIGHTS	(2) GROUP INDEX NUMBERS FOR JUNE 1936 (1923 = 100)	(3) PRODUCTS (1) × (2)
Food	33	85.6	2,825
Housing	20	77.6	1,552
Clothing	12	73.3	880
Fuel and light.....	5	84.5	422
Sundries	30	94.3	2,829
Total	100		8,508

Combined Index = $8,508 \div 100 = 85.1$

* *Cost of Living in the United States, 1914-1936* (New York: National Industrial Conference Board, Inc., 1936), p. 41.

Method of Construction

Different methods of construction are employed at various stages in the preparation of the index.

Group Indexes.—The food index is a relative of weighted aggregates. The housing index is an average of weighted relatives. The clothing index is a relative of weighted aggregates. The fuel and light index is an average of weighted relatives of the two component indexes, but the fuel component index is an unweighted average and the light component index is a weighted average. The sundries index is an average of weighted relatives of nine components which in turn employ suitable weighting systems and methods of combining individual items.⁶

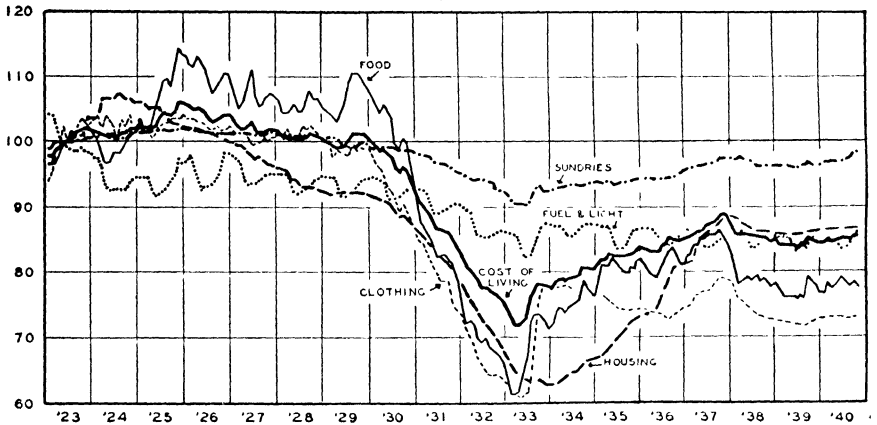
Combined Index.—The five group indexes are combined to form the Index of Cost of Living by the weighted average of relatives method. The weights, which were obtained by a method described in the section dealing with weighting, represent the importance of each expenditure group in the total family budget.

Summary

The index numbers of cost of living and its separate constituents are shown in Figure 68. The uses of such an index have multiplied greatly since 1925. Many companies, for instance, have adopted wage plans which are related to the maintenance of an acceptable level or

⁶ For more detailed explanations of these construction methods consult the original source.

FIGURE 68
NATIONAL INDUSTRIAL CONFERENCE BOARD INDEX OF COST OF LIVING,
MONTHLY, 1923-40



Reproduced from *Cost of Living in the United States, 1914-1936*, with later data added.

agreed upon standard of living for their employees. These companies claim that wages should increase as costs of living increase, and should decrease as costs of living decrease. Organized labor has been hostile to such plans because they tend to stabilize the standard of living at an existing level, and hence conflict with the long range union wage program of attaining increases and resisting decreases. In spite of this controversy, the cost of living has more and more been recognized as an important factor in attempted solutions of industrial and social problems.

Indexes of cost of living have been used in the United States in the settlement of wage disputes by arbitration boards, in the adjudication of various kinds of issues by courts of law, in the establishing of relief payments by philanthropic and governmental agencies, and in many other fields. They have also been widely accepted and used for similar purposes in other parts of the world. The International Labor Office has fostered the development of measures of living costs in various countries and has been instrumental in making them easily accessible.

AN INDEX OF INDUSTRIAL PRODUCTION

Broadly speaking, production is the creation of utilities, but common usage limits the meaning of the word to the output of material goods measured in quantitative terms. A complete index of production

would be a measure of crops grown, minerals extracted, manufactured goods produced, buildings constructed, and any other economic activities that add to an existing stock of goods. Such a broad index has never been constructed, but indexes of industrial production, based on mining and manufacturing activity, have become an important part of the statistician's equipment since World War I.

Changes in the output of manufactures and minerals are especially significant because they account for a large part of variations in the total of all economic activity and also because of the great extent to which they both affect and reflect other activities. About one-fourth of all gainfully occupied workers in this country are directly engaged in manufacturing and mining and a large part of the remainder is employed in activities that handle the materials or products of these major branches of industry. Output of manufactures and minerals is particularly important from the standpoint of the analysis of short-term movements.⁷

Measures of production are important tools for the analysis of economic and social problems and help to provide the basis for business decisions and for governmental policies. Over long periods of time they are indicators of broad changes in economic well-being; they show the results of application of human effort and technology to material resources; when broken down into component parts, they reflect shifts in patterns of living; and their short-term movements provide guides in the analysis of current business conditions.⁸

A number of indexes of industrial production have been introduced during the past twenty years. The problems and solutions involved in the construction of the one that is most widely known are described in succeeding pages.

*Index of Industrial Production in the United States,
Board of Governors of the Federal Reserve System*

In February, 1927, the Federal Reserve Board published its new "Index of Industrial Production."⁹ For four years before that the Board had been publishing a current index of production in basic industries which provided a crude measure of industrial and mineral output in the United States. The great increase in the amount of regularly collected statistical data concerning various phases of production, and the promptness of their availability, made it possible to revise the

⁷ Woodlief Thomas and Maxwell R. Conklin, "Measurement of Production," *Federal Reserve Bulletin* (September, 1940), p. 915.

⁸ *Ibid.*, pp. 912-13.

⁹ "A New Index of Industrial Production," *Federal Reserve Bulletin* (February, 1927), pp. 100-103; (March, 1927), pp. 170-77.

older index in order to make it much more comprehensive and to improve its method of construction. Again in 1940, the Board published a revision of its index "to provide a broader and more accurate measure of current changes in the physical volume of industrial output."¹⁰

Purpose

The Board of Governors of the Federal Reserve System is the policy-determining body of the banking structure of the United States. For its guidance many different kinds of statistical information are compiled. The absence of adequate current data concerning industrial production led the Board to prepare its own index. The results of the index have now become such an important part of business literature that it would not be too much to say that the index is fully as valuable to the business community as to the Board of Governors.

Explicitly the purpose of the index is to measure monthly the volume of production of mineral products and manufactured goods. This measure is expressed as a fixed base index number which permits comparison of short term or longer period changes in production. Separate indexes are published for major groups, subgroups, and individual commodities.

Kinds and Sources of Data

Requirements of Data.—The nature of this index and its general purposes impose a number of restrictions upon the selection of data to be included.

1. The need for publishing the data promptly, that is, within four or five weeks after the end of the month, with preliminary estimates earlier, greatly restricts the choice of series of data.

2. The requirements that all sources be reliable and that data reported be continuously uniform limit the sources to organizations and agencies which are well organized for data-collecting and reporting.

3. The length of the period of the index requires that it include only series which have been collected regularly each month since the beginning of 1923.

4. The development of new products and changes in qualities of old products present a problem that has been faced previously and requires careful attention to revisions and adjustments.

¹⁰ "New Federal Reserve Index of Industrial Production," *Federal Reserve Bulletin* (August, 1940), p. 753.

Sources.—The data included in the index are obtained from many types of regular reports, releases, and publications, furnished by public and private agencies such as: Bureau of the Census, Federal Trade Commission, Bureau of Foreign and Domestic Commerce, Bureau of Mines, The American Iron and Steel Institute, Iron Age, Russell's Commercial News, and the Rubber Association of America.

Groups Included.—Eighty-one individual series of data are employed in the construction of the index. They represent all principal groups of industries in manufacturing and mining at some stage in the production process, distributed among 16 groups of manufacturing industries and 2 groups of mining industries. The manufactures section of the index is divided into two parts: durable manufactures, composed of 6 major groups, each one of which is further divided into two or three subgroups; and non-durable manufactures, composed of 10 major groups each of which is also further divided into subgroups. The two major groups included in the minerals index are fuels and metals. The groups and their relative magnitudes or weights in the index are shown in Table 110.

Treatment of Data.—Prior to the latest revision, the data included in the index of industrial production were based on figures representing average output per working day. In the new index numerous revisions of these working day allowances produced noticeable changes in the daily averages. In certain cases where the necessary data are available, production per man-hour has been substituted for production per working day. The entire index may be shifted to a man-hour basis at some future date when suitable data are obtainable.

Base Period

In 1926 when the Federal Reserve Index was first constructed, the three preceding years, 1923, 1924, and 1925, were selected as the base period because the period was familiar. This three-year period also provided a broader base for comparison than a single year, and tended to reduce the extreme effects of sharp variations which might have occurred in a series in any one of the three years.

The validity of this principle is confirmed by the authors of the 1940 revision in the statement: ". . . it is desirable in an index used for current analysis to select a base period which is comparatively recent and familiar and which is not characterized by extreme variations.

TABLE 110

GROUPS OF MANUFACTURING AND MINING INDUSTRIES INCLUDED
IN THE NEW FEDERAL RESERVE INDEX OF INDUSTRIAL PRODUCTION *
Percentage Distribution of Base Period Value Weights

GROUP	PER CENT DISTRIBUTION
Industrial Production—Total	100.00
Manufactures	84.80
Durable Manufactures	37.93
Iron and steel	11.00
Machinery production	10.81
Transportation equipment	5.92
Nonferrous metals and products	2.81
Lumber and products	4.39
Stone, clay and glass products	3.00
Nondurable Manufactures	46.87
Textiles and products	11.22
Leather and products	2.28
Manufactured goods products	10.92
Alcoholic beverages	1.84
Tobacco products	1.24
Paper and products	3.13
Printing and publishing	6.44
Petroleum and coal products	2.14
Production of chemicals	6.27
Rubber products	1.39
Minerals	15.20
Fuels	13.01
Metals	2.19

* Thomas and Conklin, *op. cit.*, p. 919.

A base selected to cover a number of years of different movements may average out some of the variations likely to be reflected in too short a base period."¹¹ Accordingly the five years, 1935–39, were chosen because they seemed to meet the principal requirements of a base period. In addition various government departments and agencies, acting together through the Central Statistical Board, had agreed upon the adoption of the 1935–39 period for the construction of all index numbers unless for special reasons some other period were preferred as a base.¹²

¹¹ Thomas and Conklin, *op. cit.*, p. 920.

¹² The following statement by the Central Statistical Board sets forth the reasons for adopting the period 1935–39:

"Adoption of a 1935–39 base for all general purpose index numbers prepared by Federal agencies was recommended by the Central Statistical Board at a meeting held on May 23. The use of a uniform base period should facilitate comparison of the changes shown by such indexes. At present a multiplicity of base periods prevails. The Department of Agriculture publishes some index numbers on a pre-war base and others on a 1924–29 base; the Board of Governors of the Federal Reserve System uses a 1923–25 base; the Department of Labor, a 1923–25, a 1926, and a 1929 base; and the Department of Commerce, a 1923–25, a 1929 and a 1929–31 base.

"A more recent base has been urgently needed for index numbers for the following reasons: (1) Many statistical series are not available before 1935. Inclusion of such series in index numbers having base periods prior to that year forces fictitious adjustments of the base average. (2) The significance of any base period as point of reference depends

Method of Construction

Production of different kinds of goods is recorded in a wide variety of units, such as tons, pounds, bales, feet, etc. "The combination of these different units into a single composite is the basic technical problem of index number construction."¹³ This problem was met in the old index by computing value aggregates for individual commodities, groups of commodities, and finally for all commodities. The ratio between corresponding aggregates for a current period and for the fixed base period provided the desired index numbers.

In the revision of 1940 the method has been shifted to an average of weighted relatives. The weights are computed by a formula that uses base-period quantities and consequently produces a result that is equivalent to what would be obtained by an aggregate method. The average of relatives form is used because it simplifies the intermediate calculation steps, is better adapted to the form of the weights, and permits more comparisons between partial and total indexes.¹⁴

somewhat on the assumption that the index series may be expected to fluctuate around this level in the future. Important changes in economic relationships during recent years have largely destroyed the significance of pre-depression base periods.

"The five-year period, 1935 through 1939, is the most suitable recent period for adoption as a standard base. It is neither predominantly a period of very high business activity nor one of very low business activity. It is long enough to meet the needs of agricultural indexes. It is recent. It includes 1939, for which decennial census data will shortly be available. It also covers three censuses of manufactures; one census of agriculture; two censuses of business; and one census of electrical industries. Because of its recency, there are far more bench-mark data (in addition to those from the census) available than for any earlier period.

"It was recognized by the Board that the need for adopting a new and recent base will recur periodically, although too frequent changes in base periods are not desirable. The Board therefore recommended that the question of base periods be reexamined before the end of another decade, and that consideration then be given to shifting the standard base period forward to a more recent series of years."—"New Federal Reserve Index of Industrial Production," *Federal Reserve Bulletin* (August, 1940), p. 760.

¹³ Thomas and Conklin, *op. cit.*, p. 917.

¹⁴ Expressed in the symbols of the preceding chapter the formula is,
$$I = \frac{\sum \left(\frac{q_k}{q_0} \times q_0 p_s \right)}{\sum (q_0 p_s)},$$

in which k is any current month, o is the average of the years 1935 to 1939 and s is the year 1937. Obviously, by cancelling q_0 in numerator and denominator of the upper summation the formula

reduces to $I = \frac{\sum (q_k p_s)}{\sum (q_0 p_s)}$ which is the usual aggregative form with base period prices (unit

values) as weights. The Board has found it more convenient to use the average formula in the form

$$I = \sum \left(\frac{q_k}{q_0} \times W \right)$$
 in which $W = \frac{q_0 p_s}{\sum (q_0 p_s)}$, which amounts to using value weights in

the form of a per cent distribution instead of in dollars.

Weights

For manufactured products, the basis of the weights employed in the index is the value added by the process of manufacture. This figure, which is the total value of the product minus cost of materials, fuels, and containers, is used as a measure of importance because it is expressed in a way common to all the manufacturing series included. Furthermore, where the production process uses other manufactured products as components, value added by manufacture eliminates the duplication that would be involved if value of product was employed. In the case of several of the minor series, however, individual treatment is necessary because value-added data are not available. In weighting the mining industries, total value of product is used.

To determine the weights for each of the 16 manufacturing series the following four steps have been taken:

1. The total value added by all manufacturing industries in 1937 (the latest completed census of manufactures) is distributed among the 16 component groups in proportion to the value added by each group.
2. The value added in each group is then ascribed to the individual series included in the Federal Reserve index according to what phases of production each series is intended to represent.
3. Each individual production index for 1937 is divided by the average value of that index in the base period, 1935–39, to obtain the relative of 1937 production to base period production.
4. The value added in 1937 ascribed to each series in step 2 is then divided by the relative of step 3 to obtain hypothetical value-added figures for each series in the base period 1935–39.
5. The hypothetical 1935–39 value-added figures calculated in step 3, are expressed as a percentage distribution. These are the percentage weights defined as W in the formula in footnote 14 on page 519 and listed in Table 110.

By similar steps the weights for the minor components of each of the 16 groups are calculated. For the weights of the minerals group a similar procedure is followed using total value produced instead of value added.

Step 2 has been introduced in order to use value weights that are based on the industrial price structure in 1937, combined with the average physical measures of productivity in 1935–39. Those in charge of the construction of the new index admit that, "The implicit assump-

tion in this method is correct only if prices (or rather values added per unit) of every product show no change or change by the same proportion. Although it can be readily shown that this is not the case, the differences are likely to be small. Short-term changes in the relative importance of different industries are due mostly to changes in quantities of goods produced and only to a smaller extent to variations in unit values."¹⁵

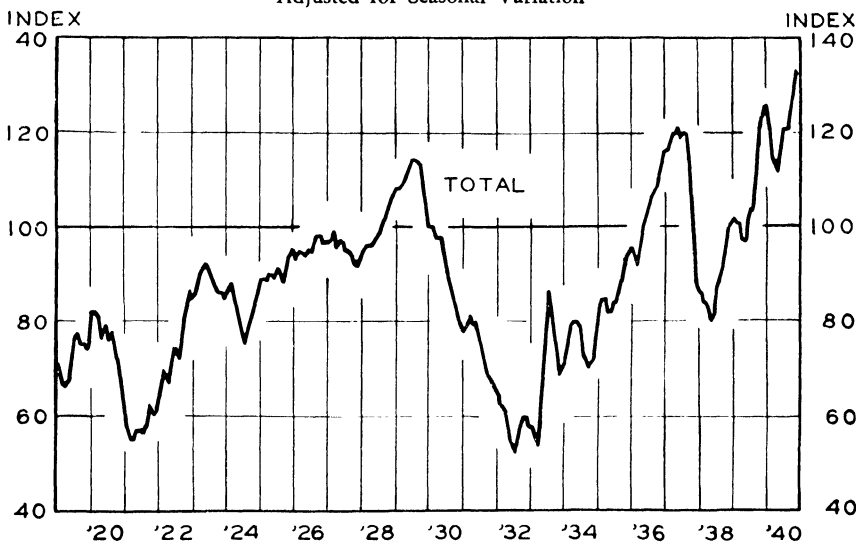
To allow for shifts in the importance of different items during the period 1919 to date, three sets of weights are used: (1) for the most recent period, 1929 to date, the weights just described are employed; (2) for the preceding period, 1923 to 1929, weights are calculated according to the same four steps, but using value-added data from the 1923 Census of Manufactures instead of 1937 data; (3) for the period 1919 to 1923, the weights are based on both 1919 and 1923 data.

Summary

The course of the Index of Industrial Production is shown in Figure 69. The upward trend during the twenties, interrupted by the depres-

FIGURE 69

FEDERAL RESERVE BOARD INDEX OF INDUSTRIAL PRODUCTION, MONTHLY, 1919-40
Adjusted for Seasonal Variation



Reproduced from *Federal Reserve Bulletin*, September, 1940.

¹⁵ *Ibid.*, p. 919.

sion of 1932, appears to have been resumed in 1940. It is believed, however, that the index understates the growth of industrial production, since there is no way of including measures of the change in quality of products. In spite of this understatement it is generally agreed that this index is a reasonably accurate indicator of both short- and long-term changes in industrial output and that it may be employed as a dependable tool in economic analysis.

AN INDEX OF EMPLOYMENT

Indexes of employment have been recognized as important administrative tools in many parts of the world. Greatest use of them has been in the highly industrialized areas. One of the most reliable indexes of this kind is that developed by the United States Bureau of Labor Statistics.

The United States Bureau of Labor Statistics Index of Employment

Purpose

Early in the development of business indicators in the United States, the Bureau of Labor Statistics determined to produce indexes of employment and payrolls. The specific purposes of these indexes were: (1) to indicate changes in the labor market, that is, the number of persons actually on payrolls, classified by line of industry and by geographic areas; (2) to show the course of manufacturing activity in the short run; and (3) to provide a rough measure of the amount and direction of the flow of purchasing power.

The detailed explanation which follows is confined to the employment index; accordingly the emphasis falls on the first of these three purposes. An explanation of the payroll index would differ from the employment index in minor details only.

Kinds and Sources of Data

The Bureau of Labor Statistics makes a direct canvass of manufacturers every month to obtain data for the index. In 1915, when this work was started, three states, Massachusetts, New York, and New Jersey, were already collecting data. It was possible for the Bureau to use the data collected by these states, but it was necessary to obtain

data from a much broader area than this to make the results representative. Direct canvass of employers seemed to be the only method available.

Working with various agencies, employers' associations, trade associations, state industrial commissions, etc., the bureau has built up a group of voluntary co-operating reporters covering over 50 per cent of the employment in manufacturing industries in the United States. Some industries are well represented, as much as 80 to 90 per cent of their employment being included in the reports. For some other industries, no more than 30 to 50 per cent of the employment is reported. Every state is now included in the sample and constant efforts are being made to increase the representation.

Base Period

The base period for the index of employment has been changed several times since its introduction. At first, the average for the year 1919 was selected; subsequently both 1923 and 1926 were used; and at the present time, the three-year average 1923-25 is employed. For a few industrial groups which were added later, particularly those of the non-manufacturing classifications, data were not available until a later time. Consequently, for these groups the base period is the first year of the new collection of data, usually 1929.

Two main reasons justify the use of the three-year average, 1923-25, as the base for the index. In the first place, this three-year period was one of relative stability in employment. Secondly, the period is one which has been used as a base by so many other governmental and private agencies in constructing indicators of economic and business conditions that it was thought advisable to employ it in this index. Greater range of comparability is the result.

Weights

Weights are employed in this index only for combining the minor group indexes into industry indexes and the latter into a composite index representing all employment in manufacturing. The weights are introduced in order to insure that each group and industry is given the same relative importance in the total index that it has in employing workers. This weighting system is parallel to that explained on pages 393-95 in describing the construction of an index of independent

retail trade in Ohio. The purpose of the weights is to adjust a sample with variable representativeness in its several parts to conform to a universe in which the relative importance of the several parts is known.

The weights are taken from the Census of Manufactures using the average number of workers in each industry as determined from the average of the censuses of 1923 and 1925 and an estimate for 1924. Table 111 gives the weights for the major industry groups.

The index of employment is computed by multiplying the index of each of the industry groups shown in the table by the average number of wage-earners in that industry in 1923 to 1925, adding these products together, and dividing the sum of these products by the sum of the weights. The result is an index covering total employment in manufacturing industries.

TABLE 111

WEIGHT FACTORS USED FOR COMPUTING INDEXES FOR MAJOR INDUSTRIAL GROUPS IN
BUREAU OF LABOR STATISTICS INDEX OF EMPLOYMENT *

INDUSTRIAL GROUP	AVERAGE NUMBER OF WAGE EARNERS (Census 1923 and 1925, and Estimates for 1924)
Total Manufacturing	8,381,700†
Iron and steel and their products, not including machinery.....	859,100
Machinery, not including transportation equipment.....	878,100
Transportation equipment	563,500
Railroad repair shops	482,100
Non-ferrous metals and their products.....	282,600
Lumber and allied products.....	918,400
Stone, clay, and glass products.....	350,300
Textiles and their products.....	1,629,400
Fabrics	1,105,600
Wearing apparel	474,100
Leather and its manufactures.....	323,500
Food and kindred products.....	668,300
Tobacco manufactures	138,400
Paper and printing	531,100
Chemicals and petroleum products.....	333,000
Other than petroleum refining.....	268,200
Petroleum refining	64,800
Rubber products	134,300

* Lewis E. Tolbert and Alice Olenin, *Revised Indexes of Factory Employment and Payrolls, 1919-1933*, Bull. 610 (United States Bureau of Labor Statistics), pp. 9-10.

† Includes miscellaneous industries not shown in the subgroups.

Method of Construction

The nature of the data and the methods of their collection for this index determine the construction method to be used. The data themselves represent the number of employees on the payrolls of reporting firms at the payroll date nearest the 15th of the month. The data are collected from a vast number of firms all over the United States and

represent about 50 per cent of factory workers. Due to many circumstances the firms that report every month are not identical: some of them go out of business, others are delayed in returning their reports, and still others are kept from sending in their reports by a variety of circumstances peculiar to conditions in the different industries or sections of the country. A method of construction must be used which considers these two factors: (1) large numbers of reports; and (2) differing numbers of reports from month to month.

The chained relative of weighted aggregates meets these requirements. Each month, as reports from co-operating firms are received, they are posted on individual firm records of employment. When a sufficiently large number of reports has been received or the agreed-upon date for computing the index is reached, the records of all firms that have employment figures posted for the latest two months are separated according to industrial classifications. For each classification, the numbers of persons employed in the latest two months are added. These two totals, representing the numbers of persons employed on the 15th of the latest two months in identical establishments, are then expressed as a relative, i.e., the total of the latest month is divided by the total of the former month. This quotient is then multiplied by the index of the preceding month to obtain the new index. The following example illustrates the method of computing:

Total number persons employed in 629 establishments, June 1939 = 12,576
Total number persons employed in 629 identical establishments, July 1939 = 12,342

$$\text{Relative, } \frac{\text{July}}{\text{June}} = \frac{12,342}{12,576} = .9814$$

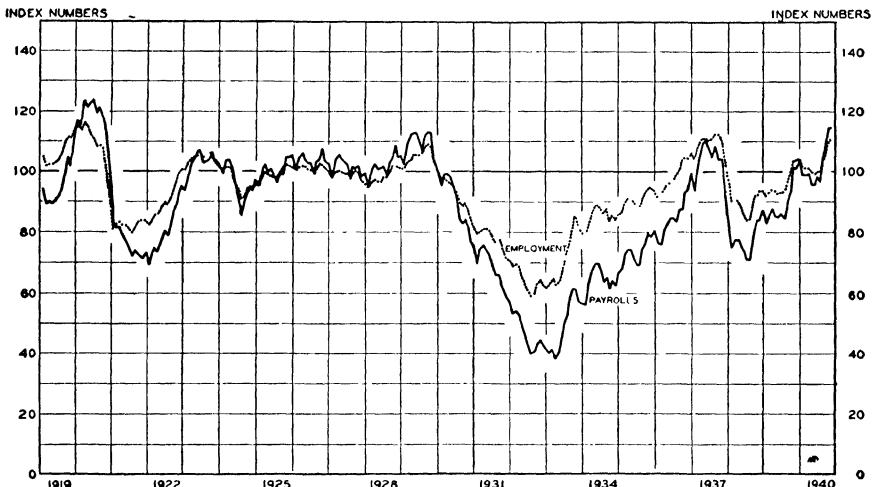
Index for June..... 95.63
Index for July = .9814 $\times$ 95.63 93.85

This method is a chaining of a monthly relative of aggregates to the previously derived index, which started as a relative of aggregates in the base period.

Since the index does not accurately reflect "the rise in employment occasioned by the entrance of new firms into industry, nor the decline of employment due to industrial mortality,"¹⁶ the Bureau of Labor Statistics revises its indexes every two years to the levels of employment as established by the biennial Census of Manufactures. This adjustment has in every case tended to increase the index.

¹⁶ Tolbert and Olenin, *op. cit.*, p. 11.

FIGURE 70
INDEXES OF EMPLOYMENT AND PAY ROLLS, MONTHLY, 1919-40
 Unadjusted for Seasonal Variation



Data published by the United States Bureau of Labor Statistics.

Summary

Changes in the size of the labor force provide a sensitive and easily obtained indicator of changes in the business situation from several points of view. To working men themselves, changes in employment indicate the condition of the labor market. To employers, changes in employment signify fluctuations in production, but in addition they may point to failure in planning and organization. To government agencies, especially those engaged in the new fields of social security, changes in employment serve as guides to administrative decisions. The fluctuations in the employment index from 1919 to 1940 are shown in Figure 70, in comparison with the payrolls index.

A WHOLESALE PRICE INDEX

Prices have constituted the central and most perplexing problem of political economists for centuries. Prices are related to every phase of economy—social, governmental, or business. A nation's international trade, its relations with other nations, its holdings and use of resources including the precious metals, and the welfare of its people—all have been expressed in terms of prices.

Some of the difficulties encountered in the past in the collection of reliable price data have been the acceptance of definitions and standards for commodities, the establishing of comparable levels of trade, the maintenance of regular reporting, and the securing of current information. These collection problems persist today but additional experience of statisticians and better co-operation by the business community have combined to reduce their importance. The most reliable prices originate in wholesale markets where buyer and seller have nearly equal bargaining power and regular recording of transactions is an integral part of the marketing machinery. Further, wholesale prices are fundamental in our system of production and distribution.

Index numbers of wholesale prices have accordingly attained the maximum development and the widest use. Indexes of retail prices, wage rates, construction costs, and other price phenomena have been developed since World War I, but the major emphasis remains on the wholesale price index. The description which follows illustrates the solution to the problems of construction arrived at by a trade association in preparing an index of wholesale prices for the use of its members.¹⁷

*Weekly Wholesale Price Index of the
National Fertilizer Association*

Purpose

As the name implies the general purpose of this index is to measure the movements of commodity prices. To that end it must be broad enough in scope to be representative. This means that quotations must be included for different kinds of markets and for varieties of commodities at different stages of production. The fact that the index is prepared and published weekly¹⁸ implies that current figures will be made available as soon as possible after the period they cover.

To be useful to fertilizer manufacturers the index has in addition several special features. Two of the eleven groups of commodities in the index consist of prices of fertilizer raw materials and finished

¹⁷ Herbert Willett, *The Weekly Wholesale Price Index* (3d ed., Washington: The National Fertilizer Association, February, 1941).

¹⁸ "Based largely on closing market prices on Thursday, Friday, and Saturday, the index is released to the press on Monday forenoon of the following week. It is widely quoted in business periodicals [e.g., the *Commercial and Financial Chronicle*] and in daily newspapers from coast to coast." *Ibid.*, p. 5.

products. Prices in the farm products group are emphasized because fluctuations in farm income are directly related to the quantity of fertilizer that can be sold.

Kinds and Sources of Data

Prices Used.—The all-commodity price index is based on 250 separate price series classified in eleven major groups, and separate index numbers are computed for each group. The Association points out that the coverage is really broader than is indicated by 250 series, because some of these series are averages or composites of a number of quotations.

The list of commodities included in the index is necessarily restricted by the requirement of prompt publication. Only those commodities are used for which prices are currently available on a continuous and comparable basis. This limitation excludes only a few important commodities such as tobacco for which representative and accurate price series are very difficult to obtain. Whenever a great many quotations for certain commodities are available, as is the case with wheat, prices have been selected only for the most important kinds and grades or from the leading markets.

Sources.—Wholesale prices in this index include quotations on raw materials, semi-finished goods and finished goods but exclude prices involved in sales to ultimate consumers. Prices of 168 commodities, two-thirds of the total of 250, are compiled regularly from trade and business periodicals. Prices of 25 fertilizer items are collected by the Association directly from its members. Prices of 40 farm machinery products are United States Bureau of Labor Statistics data collected directly from manufacturers. The remaining 17 price series are also provided by governmental agencies.

Commodities.—The price series included in the index are classified according to industrial groups with subdivisions in that particular area of production and trade which the index is especially designed to serve. For instance, fertilizer materials, fertilizers, and farm machinery are listed as separate groups because of their importance to the fertilizer industry.

Fertilizer materials include 27 price series, collected once each month. Because of the frequent change in materials used for making fertilizers, the latest available data are used in determining the weights. The fer-

tilizers group includes prices of 25 commodities based on quotations from all fertilizer-consuming areas, representing about half of the commercial fertilizers sold in the United States. Prices of these commodities are collected once each month as of the 15th. The index of farm machinery prices is computed monthly by the United States Bureau of Labor Statistics and is shifted by the Association to a 1935-39 base. It is based on prices of 40 items of farm machinery with from two to eight quotations for each item collected from 31 manufacturers. Since indexes for these three groups are computed but once a month, they remain constant for four or five weeks in the weekly all-commodities index.

Seven of the principal groups in the index are similar to those included in other prominent indexes. Foods with 45 price series is the largest group in the index. All important food items excluding fresh vegetables and fruits (except bananas and oranges) are represented in the group. A separate weekly index is computed for the fats and oils and cottonseed oil subgroups. Prices of milk and eggs appear in the farm products group and in practically unchanged form in the foods group; they appear only once in the all-commodity index, however. The textile group is based on 26 price series of raw materials and semi-finished goods.

Raw cotton prices are included in the textile group with a weight based on domestic consumption of cotton, and in the farm products group with a weight based on total cotton marketings. The fuels price index is computed from nine series, including anthracite and bituminous coal, coke, crude petroleum, fuel oil, gasoline, and kerosene. The metals group includes price series of thirteen raw materials and semi-manufactured products. Three of the series represent composite prices which are averages of a large number of quotations for finished steel, pig iron, and steel scrap as compiled weekly by *Iron Age*. Fifteen price series with lumber receiving the greatest weight are included in the building materials group. The chemicals and drugs index is computed from 20 price series of which industrial chemicals are the most important. Prices of commodities which do not seem logically to fit in the foregoing groups are put in a miscellaneous class. The series included in the group are: hides (2 series); leather; calfskins; rubber, crude; lubricating oil; wood pulp; news roll paper; book paper; paper board; cottonseed meal; linseed meal; bran; middlings; tires; and cigarettes.

Base Period

In January, 1941, the base for this index was changed to the five-year period, 1935–39, which conforms with the base period recommended for index numbers compiled by governmental agencies (see footnote 12, pages 518 and 519, in this chapter).

Weights

In the selection of quantity weights used as multipliers for each price series included in the construction of the index, consideration is given to the relative importance of the commodity in question. . . .

Since in some cases the prices of individual commodities are used as samples to represent a group of commodities the weights used must be increased so that the entire group will be adequately represented. . . .

Changes are made from time to time in the quantities as changing conditions shift the relative importance of various commodities. Use is made of recent representative periods when conditions were comparatively stable. As changes are made the revisions are applied to the compilation of aggregate values for the base period as well as to current values.¹⁹

Method of Construction

This index is of the relative of weighted aggregates type. For each commodity included in the index, the average price during the period 1935–39 is multiplied by the quantity weight assigned to that commodity. The product is the value of that commodity in the base period. These individual values are combined into groups for group indexes and are also added together to obtain the base value for all commodities. Each week as each price is received, it is multiplied by the corresponding base period quantity to give the current weekly value of that quantity. These values are then added to provide group totals and the all-commodity total, and each is divided by the comparable base period value to obtain the group and all-commodity index numbers.

Expressed in the usual symbols the computation is $\frac{\sum(p_k q_r)}{\sum(p_o q_r)}$ in which k = any current week; o = the period 1935–39; r = the period for which the weight of each commodity is used.

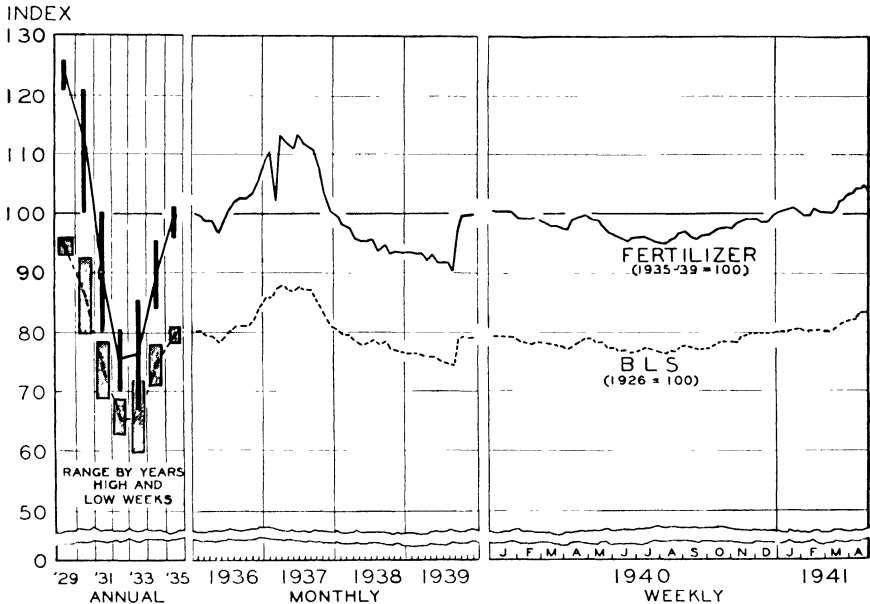
Summary

The course of this wholesale price index since 1929 can be followed in Figure 71 in comparison with the Wholesale Price Index of the

¹⁹ *Ibid.*, p. 10.

FIGURE 71

WHOLESALE PRICE INDEXES OF NATIONAL FERTILIZER ASSOCIATION AND UNITED STATES
BUREAU OF LABOR STATISTICS; ANNUALLY, 1929-35; MONTHLY, 1936-39;
WEEKLY, JANUARY 1940-APRIL 1941



Data published by National Fertilizer Association and United States Bureau of Labor Statistics.

Bureau of Labor Statistics. The latter is on a 1926 base, so that the extent of variation of the two curves is not strictly comparable. However, it is apparent that the movements of the two indexes have coincided quite closely throughout the entire period. The Fertilizer Association Index has been more sensitive to price changes, particularly during the last year.

A LOCAL BUSINESS INDEX

Business indicators for cities and industrial areas have been very much in demand in the past several years. Chambers of Commerce, local clubs of businessmen, and business executives responsible for the policies of their concerns have steadily insisted on having more indicators of local business. These demands have led to the compilation of a great deal of data and the calculation of a variety of indicators for local use. In a few instances, enough suitable data have been

available to make possible the computation of composite barometers of business.²⁰

Except in the largest cities, however, the number of series available as business indicators is usually rather limited. In many cases the only series that is easily and quickly obtainable is bank debits,²¹ which is one of the series usually included in composite barometers. The use of this series will be illustrated by a description of the index constructed for Canton, Ohio.

Index of Bank Debits in Canton, Ohio

Purpose

Bank debits are the total charges of checks to individual accounts.²² As the volume of business increases, more and larger checks are written to meet payrolls and to pay for goods and expanding business facilities. A decrease in business, conversely, is quickly reflected through a reduction in the number and value of checks written.

There are certain disadvantages that must be kept in mind when using bank debits as a general indicator of business. In the years since 1930, a number of restrictions have been placed on the use of bank checking accounts. Banks have instituted service charges in such a way as to reduce the number of smaller accounts and the number of checks written by the users of large accounts. For a short period of time the federal government levied a tax on each check written. These practices have resulted mainly in reducing the number of checks drawn for the payment of consumption goods although they have

²⁰ See chapter XXV for a presentation of this subject.

²¹ The Board of Governors of the Federal Reserve System has been publishing weekly totals of bank debits (and many other banking series) since 1919. Reports of these data for many cities in the United States can be obtained upon request to the Board, Washington, D. C.

²² Debits to individual accounts include the charges made on the books of reporting banks to all deposits except interbank deposits. They cover debits to the time and demand deposit accounts of individuals, partnerships, corporations, and governmental units; included are debits to postal savings accounts, other savings accounts, payments from trust accounts on deposit in the banking department, and certificates of deposits paid; excluded are debits to the accounts of other banks or in settlement of clearing house balances, payments of certified and officers' checks, charges to expense and miscellaneous accounts, corrections, and similar charges. Looked at from another point of view, debits to individual accounts comprise check payments (1) for goods in various stages of production and distribution, (2) for services, i.e., wages, salaries, rents, dividends, taxes, etc., (3) in financial transactions, such as property transfers and security trading, and (4) mere transfers of funds, as in gifts, in making and repaying loans, and in shifts of deposits between accounts. Because it is so all-inclusive, the term "debits to individual accounts" represents a large amount of duplication that, together with the fact that no account is taken of currency payments, limits its usefulness as a measure of business activity.

not seriously affected the total dollar volume of bank debits. In spite of these changes in practice, bank debits constitute a usable series, especially when there are no other local data readily available.

Sources of Data

The total weekly amounts of bank debits in Canton are collected after the close of business each Wednesday by the Canton Clearing House Association. Since the banks which report to the Clearing House are the largest in Canton, and do most of the banking business in the city, their total debits adequately represent the total debits to all bank accounts in the city. No problem of definition arises, nor of the collection of data, for the terms are rigidly defined by use and the Clearing House Association member banks have become accustomed to these voluntary weekly reports.

Base Period

In this case one of the most important considerations in the choice of a base year was the fact that a number of indexes for Canton have a 1926 base, e.g., Index of Building Occupancy, Index of Employment, Index of Value of Construction Contracts Awarded, and others. As a result, the monthly average in 1926 was chosen because it provided the greatest degree of comparability.

Method of Construction

Since the data in this case are reported weekly and a monthly index is desired, there are two procedures that may be followed: (1) monthly totals may be estimated by prorating the first and last weekly totals of each month,²³ (2) monthly totals may be considered as the sums of either four or five weekly totals, i.e., for each year the total for February might be considered as four weeks and the total for March as five weeks, etc. Each of these procedures had disadvantages, but since fewer difficulties appear in the second, it was the method selected.

The construction method as illustrated in Table 112 is that of a simple index. The only complication in actual calculation procedures in this case is due to the varying numbers of weeks in the different

²³ The Board of Governors of the Federal Reserve System now publishes a release showing total monthly bank debits, which are obtained by prorating the first and last weekly totals in each month.

months. Instead of using the 1926 monthly average as a base, it is necessary to compute the 1926 weekly average. Five times this average is the base for five-week months, and four times the average is the base for four-week months. In Table 112 the division process is simplified by the use of reciprocals.

TABLE 112
COMPUTATION OF INDEX OF BANK DEBITS IN CANTON, OHIO
FOR BASE YEAR AND FOR 1938*

MONTH	(1) NUMBER OF WEEKS IN MONTH	1926 (BASE YEAR)		1938	
		(2) TOTAL BANK DEBITS (thousands omitted)	(3) INDEX COL. 2 $\times$ 4 or 5 WEEKS' RECIPROCAL	(4) TOTAL BANK DEBITS (thousands omitted)	(5) INDEX COL. 4 $\times$ 4 or 5 WEEKS' RECIPROCAL
January	5	\$62,927	114.79	\$37,585	68.56
February ...	4	40,126	91.50	25,368	57.85
March	4	44,104	100.57	25,896	59.05
April	4	52,572	119.88	29,140	66.45
May	5	51,742	94.39	32,225	58.79
June	4	44,492	101.45	26,640	60.75
July	5	55,355	100.98	33,232	60.62
August	4	40,833	93.11	23,795	54.26
September ..	4	42,881	97.78	28,405	64.77
October	5	50,401	91.94	36,581	66.73
November ..	4	40,446	92.23	29,183	66.55
December ..	4	44,229	100.85	32,395	73.87
Totals ...	52	\$570,108	1,199.47		
Weekly average = Total $\div$ 52 = \$10,964			100.0		

4-week total = Average $\times$ 4 = \$43,854

5-week total = Average $\times$ 5 = \$54,818

Reciprocal of 4-week total $\times$ 100 = .0022803

Reciprocal of 5-week total $\times$ 100 = .0018242

* The Ohio State University Bureau of Business Research.

Just as the indexes for the months in the base year, 1926, were computed, the indexes for every month until the present are computed as soon as the four- or five-week totals become available. The graph of this index, Figure 72, presents a description of the movement of business in Canton from 1926-40 as measured by debits to individual accounts. A similar index might be prepared for any of the cities for which bank debits are reported.

PROBLEMS

1. State the good and bad features of each of the three methods of obtaining data concerning the cost of living.

FIGURE 72

INDEX OF BANK DEBITS IN CANTON, OHIO, MONTHLY, 1926-40

*The Ohio State University, Bureau of Business Research.*

2. Is a cost of living index a measure of prices, quantities, or values?
3. a) From Figure 68, page 514, describe the general movements of the cost of living during the years 1923 to 1940.
 b) Which of the group indexes has corresponded best with the combined index? Which one has diverged most?
 c) How could a measure of dispersion be used to obtain a precise answer to (b)?
4. What difference should there be in the types of data included in an index of production and an index of industrial production?
5. State briefly the reasons for selecting the period 1935-39 as the base of the revised Federal Reserve Index of Industrial Production.
6. Why should total value of products be used in the minerals section of the Index of Industrial Production and value added by manufacture in the manufactures section?
7. Why is it desirable in the Bureau of Labor Statistics Index of Employment to use an average of relatives in combining the group indexes to obtain the total index?
8. What is the purpose of adjusting the Index of Employment to the Census of Manufactures?

9. From Figure 70, Page 526, explain the differences in the curves of payrolls and employment.
10. Compare the two indexes shown in Figure 71. Discuss the differences in method of construction of the two indexes which might account for the relative differences in fluctuation between the two at various periods.
11. What are the strong points and shortcomings of bank debits as a measure of business conditions?

REFERENCES

DAVENPORT, DONALD H., and SCOTT, FRANCES V., *An Index to Business Indices*. Chicago: Richard D. Irwin, Inc., 1937.

Contains a very complete list of references to indexes of various kinds.

FISHER, IRVING, "A Weekly Index Number of Wholesale Prices," *Journal of The American Statistical Association*, Vol. XVIII (September, 1923).

THOMPSON, DONALD S., "Regional Indexes of Trade," *Journal of the American Statistical Association*, Vol. XXIV (March, 1929, Supplement).

CHAPTER XXI

ANALYSIS OF TIME SERIES

A CHANGE IN EMPHASIS

IN PRESENTING the principles of classification in chapter VIII provision was made for the separation of a set of data into classes according to the time of occurrence, according to the place of occurrence, or according to an attribute. The techniques developed up to this point have been generally applicable to all three types of classification. The whole purpose of this and succeeding chapters is the development of methods of analyzing the changes that occur in a series of data with the passing of time. In such analyses certain adjustments must be made in old techniques and some new techniques must be introduced.

It is not sufficient for business men to know merely that a series increased or decreased over a period of years. There are various factors at work, the net effect of which has been an increase or a decrease in the series, as the case may be. Analysis of the data involves segregation of these factors so that their separate importance can be understood. The first necessity then is to know what factors are present in a time series.

COMPONENTS OF TIME SERIES

A time series in its forward movement follows a certain course that represents the net effect of the interaction of several forces pulling it up or down. If these forces were in a state of equilibrium, the value of the series would remain constant. But in a dynamic business world the forces are never in a state of equilibrium; hence the values of time series are continually increasing or decreasing. The problems to be studied are: (1) What are the forces or components the net effect of whose interaction is expressed by the movements of a time series? (2) How can the effect of each force be measured? The first of these problems will be dealt with in this chapter, the second is deferred to succeeding chapters.

An outline of the components present in a time series follows:

A. Growth or trend

B. Rhythms

1. Cyclical fluctuation
2. Seasonal variation
3. Other rhythmic movements

C. Irregular factors

1. Determinate—Irregularity of the calendar
2. Indeterminate
 - a) Weather disasters
 - b) Labor strife
 - c) War

Growth or Trend

Trend is the tendency of data in a series to increase or decrease during a long period of time. The expression "a long period of time" cannot be defined precisely. In some cases it may mean a period as short as ten years, in others a period of one hundred years or longer. The period to which a trend can be fitted is usually limited by the data that are available rather than by any arbitrary criterion establishing the meaning of "a long period of time." In some cases the purpose of a trend determines the period to be used. Thus we speak of the post-war trend of prices, the trend of production since 1900, or the growth of the industrial structure during the past century.

There are a number of circumstances in the development of the United States which have resulted in an increasing trend in most of our business series. The growth of population, the increasing industrialization of the country, and the rising standard of living have combined to cause expansion of the production of both durable and consumable goods, of services rendered by marketing and transportation agencies and financial institutions, as well as of personal services. Series such as pig-iron production, wheat production, cold-storage holdings, freight car loadings, bank deposits, and hotel business all have increased with the expansion of the country.

Although this general relation of the growth component to the expansion of the country can be established readily, it does not follow that there is one rate of growth which is applicable to all series of data or even to those series which measure fundamental factors in our economic system. It is necessary to study the growth component of each series separately. Some will be found to have positive components,

others negative; some will change rapidly, others slowly. Most of them will vary in rate of growth over a long period of years.

There was a period in the development of statistical methods in which trend was considered as a component which could be measured by a mathematical equation of one sort or another. Most frequently a straight line was used. The dependence upon rigidly defined functions for the measurement of trend has proved to be unjustified since 1929, so that at the present time there is little agreement as to the methods of measurement and projection of trend.

The existing uncertainty concerning the measurement of trend has repercussions beyond the direct problem of trend location. The usual methods of time-series analysis employ the trend as a normal or average level of the data from which to measure and evaluate other components of a series. Hence an improper location of trend is likely to affect adversely all of the subsequent analysis.

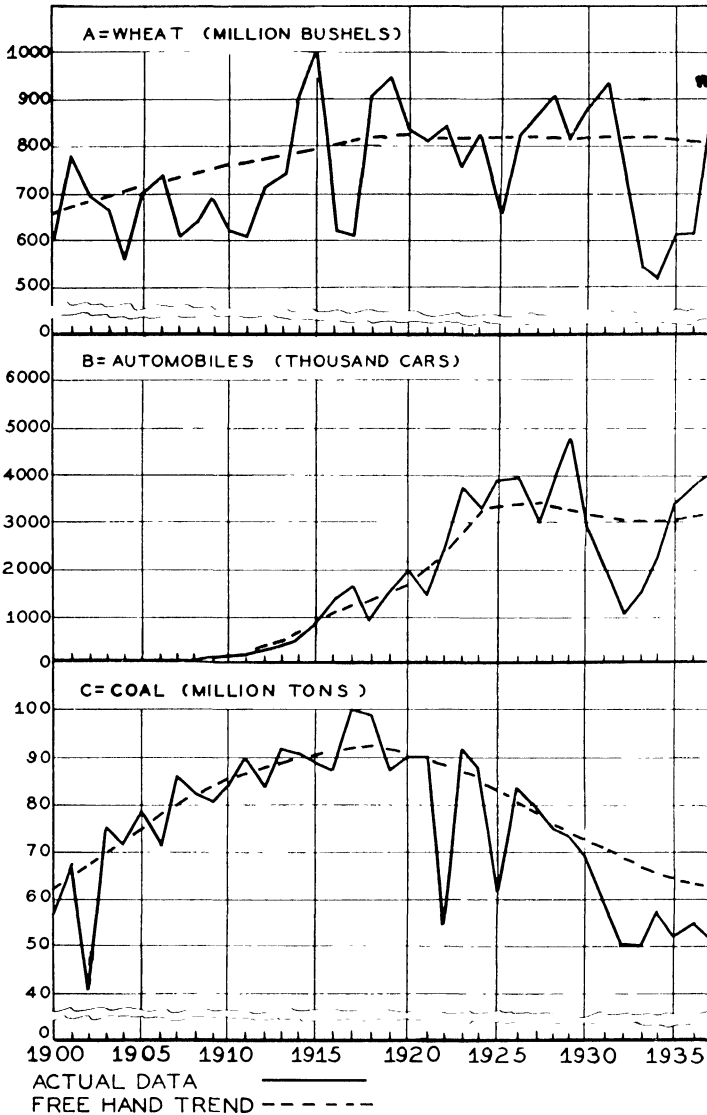
The variations in the direction and intensity of the trend component can be seen in the three curves of Figure 73 which have been plotted from the data of Table 113. All of them are fundamental series in our productive system, yet each has had a different type of growth during the 38-year period. Free-hand trends have been drawn through the respective curves to indicate approximately the paths of their growth components. Wheat production has a slowly rising trend to about 1918, but from that time until 1932 there is very little tendency toward either increase or decrease. Since 1932 the combination of drought and government control has decreased the yield markedly. Whether or not the production level of the years 1933-36 marks a new average or trend will be disclosed by what happens in future years.

Production of passenger automobiles presents a quite different trend component. The 38-year period includes the entire growth of the industry to date. From 1900 to 1923 the trend corresponds to a compound interest curve, i.e., growth at a constant rate. Since 1923 the component is nearly horizontal.

Anthracite coal production has a trend with shape quite different from either of the preceding ones. The curve increased steadily from the turn of the century to the war period but has declined since then. The marked curtailments of production in 1902, 1922, and 1925 are the results of prolonged strikes. These irregularities have been disregarded in indicating the position of the trend. It is quite possible

FIGURE 73

PRODUCTION OF WHEAT, PASSENGER AUTOMOBILES AND ANTHRACITE COAL IN THE UNITED STATES 1900-1937, AND FREE HAND TREND FOR EACH SERIES



Data from Table 113.

that the trend line is too high for 1930 to 1937. If anthracite production continues in future years at about the levels of recent years, a line much lower than that drawn on the chart will be necessary to represent the trend.

These examples are sufficient evidence of the fact that trend must be considered individually for each series. There is no such thing as

TABLE 113
PRODUCTION OF WHEAT, PASSENGER AUTOMOBILES AND ANTHRACITE COAL
IN THE UNITED STATES, 1900-1937

YEAR	WHEAT * (million bushels)	PASSENGER AUTOMOBILES † (thousand cars)	ANTHRACITE COAL * (million tons)
1900	599	4	57
1901	763	7	67
1902	687	9	41
1903	663	11	75
1904	556	22	73
1905	706	25	78
1906	741	34	71
1907	629	43	86
1908	643	64	83
1909	684	128	81
1910	625	181	84
1911	618	199	90
1912	730	356	84
1913	751	462	92
1914	897	544	91
1915	1009	896	89
1916	635	1526	88
1917	620	1746	100
1918	904	943	99
1919	952	1658	88
1920	843	1906	90
1921	819	1518	90
1922	847	2369	55
1923	759	3754	93
1924	842	3304	88
1925	669	3871	62
1926	832	3949	84
1927	875	3083	80
1928	914	4012	75
1929	823	4795	74
1930	886	2910	69
1931	942	2038	60
1932	757	1186	50
1933	552	1627	50
1934	526	2271	57
1935	626	3388	52
1936	627	3798	55
1937	874	4069	52

* *Statistical Abstract*, (U. S. Department of Commerce, 1938).

† *Automobile Facts and Figures*, (New York: Automobile Manufacturers Association, Inc., 1938).

a single trend which may be applied to a number of series. There is the further fact that the fitting of trend requires a knowledge of the background of the series of data under consideration. Perhaps it would not be too much to say that for the statistician working with time series

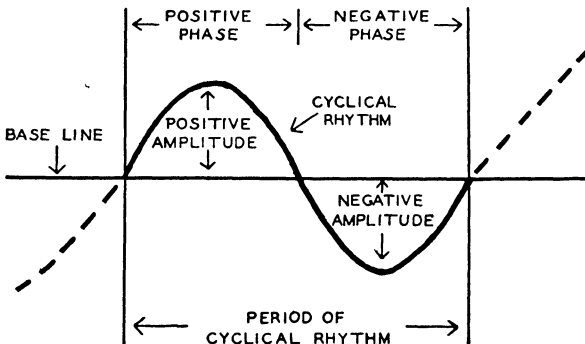
the most essential adjunct to a knowledge of statistical methods is familiarity with business history.

Rhythms

Distinct from the growth component are those movements in a time series which tend to repeat themselves over a period of time. No matter what the direction of the trend, there are always forces at work which prevent the smooth flow of the series in that direction. Various circumstances of business operations cause the series to be sometimes above and sometimes below its trend component. The most important of these circumstances are rhythmic movements which are fundamental accompaniments of our business structure. The forces causing different rhythmic movements are largely unrelated and the movements themselves can be distinguished most easily in terms of the frequency of their occurrence.

Cyclical Rhythm.—The one rhythmic movement which affects every business series is the cycle. It is a rhythm of varying amplitude and period,¹ usually referred to as the alternation of prosperity and depression. Sometimes the depression phase of a cycle has continued for five or six years, at other times for only a few months and the same is true of the prosperity phase. Likewise some cycles are extremely mild, others severe. A study of business-cycle history reveals some tendency for a severe or prolonged depression to follow a high level of prosperity, but this phenomenon is by no means universal in occurrence. In fact the only general statement that can be made concerning the relation between the phases of cycles is that depression has succeeded

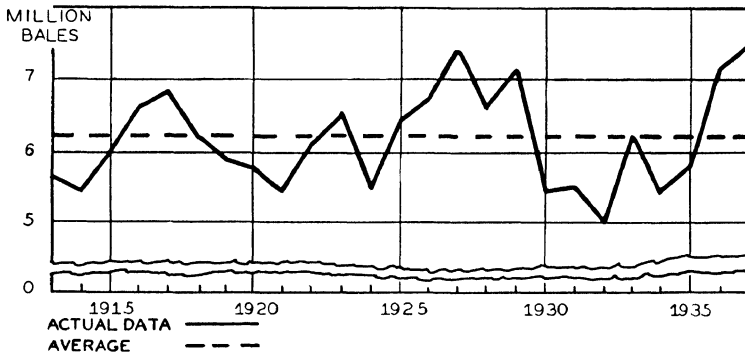
¹ The terms "amplitude," "period," and "phase" as used in describing business cycles can be defined best by reference to a diagram.



prosperity and prosperity has succeeded depression with considerable regularity. Therefore we speak of the cyclical rhythm. The path followed by the cyclical rhythm can be seen in Figure 74 containing the annual mill consumption of raw cotton. This series serves particularly well since it contains very little long-time growth. Even a cursory examination of the figure shows that the cycles follow no regular pattern, and without some background knowledge of general business conditions it would be difficult to select the periods of the cyclical movements.

Three complete cyclical swings and part of a fourth are shown on

FIGURE 74
CONSUMPTION OF RAW COTTON IN THE UNITED STATES 1913-1937



Data from Table 114.

TABLE 114
CONSUMPTION OF RAW COTTON BY MILLS IN THE UNITED STATES, 1913-37
(Thousand Bales)

YEAR	COTTON CONSUMED *	YEAR	COTTON CONSUMED *
1913	5,583	1926	6,684
1914	5,449	1927	7,407
1915	6,009	1928	6,570
1916	6,620	1929	7,051
1917	6,816	1930	5,377
1918	6,177	1931	5,450
1919	5,920	1932	5,016
1920	5,843	1933	6,210
1921	5,407	1934	5,419
1922	6,088	1935	5,650
1923	6,522	1936	7,104
1924	5,521	1937	7,425
1925	6,432	Average	6,148

* 1913-22: "Record Book of Business Statistics," *Survey of Current Business* (1927), p. 20. 1923-37: *Survey of Current Business* (Annual Supplements, 1932, 1936, and 1938), pp. 260, 141, 154, respectively.

the chart. The first runs from a low point in 1914 to a high in 1917 and back to a low in 1921. Hence this cycle has a period of seven years. The second cycle runs from a low point in 1921 to a high in 1923 and back to a low in 1924, and the period is three years. The third cycle is eight years in length, running from 1924 to 1927 to 1932. Since 1932 the curve has been increasing, but the high point is indeterminate as this is being written. The temporary rise in 1933 is not considered to be cyclical in character but rather an accompaniment of the National Industrial Recovery Act.

The difference in amplitude from cycle to cycle can be seen in the list below. The variations of cotton consumption from the average of the 24-year period at the high and low points are expressed as per cents of that average.

	PER CENT
1. Low—1914	-11.4
2. High—1917	+10.9
3. Low—1921	-12.1
4. High—1923	+ 6.1
5. Low—1924	-10.2
6. High—1927	+20.5
7. Low—1932	-18.4

The tendency toward small amplitude accompanying a short cycle and large amplitude with a long cycle can be seen in the 1921–24 cycle and in the 1924–32 cycle. It would be unwise, however, to take this relationship as a general rule, because many examples of short-period, high-amplitude cycles, as well as long-period, low-amplitude cycles will be found in practice.

Particular mention should be made of what is known as the *price cycle*. Series that are expressed in dollars, e.g., bank debits, sales, and income tax collections, contain a source of cyclical fluctuation not present in series expressed in other units. That is, since prices themselves tend to increase and decrease cyclically, a series such as sales of grocery stores decreases more in depression and increases more in prosperity than would result from the actual changes in the amount of goods handled. The cyclical nature of prices themselves can be studied by the use of price indexes. But in dealing with grocery store sales or any similar dollar series the object is usually to separate the several components. The price component can be separated from the residual cyclical component by methods which are explained later in this chapter. The methods of deflating a series for price changes are also discussed in chapter XIX, pages 495–97. This step then

is ordinarily a part of the analysis of series expressed in dollars, and becomes an indispensable part if the given series is to be compared with others not expressed in the dollar unit.

There are few general rules which can be established concerning the cyclical rhythm. It varies in both amplitude and period. There may be several minor fluctuations within a single cycle. In fact there is not even unanimous agreement that the cycle exists. Some writers prefer to consider all of the movements of such a curve as that shown in Figure 74 to be irregular. They admit, of course, the response of the curve to the alternation of prosperity and depression, but deny the existence of sufficient regularity in the movements to justify the application of the notion of rhythm. Nevertheless the succeeding discussion is based on the assumption that the cyclical rhythm exists, however faltering its course.

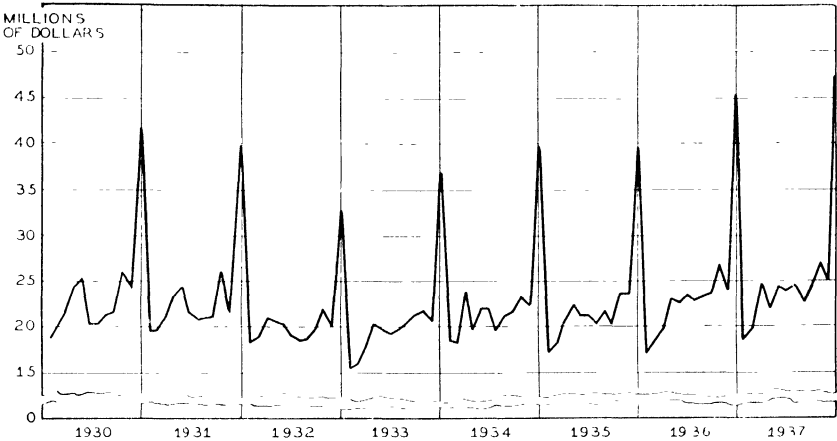
Seasonal Rhythm.—The word "seasonal" implies that we are dealing with something related to the weather. This is partly but not entirely true. Seasonal variations are of two kinds: (1) those resulting from natural forces; (2) those resulting from man-made conventions. Some examples will clarify this distinction. In the northern part of the United States construction work of all kinds is greatly curtailed during the winter season. Hence data concerning road construction, building activity and the like have seasonal variations that are directly related to the weather. The same thing is true of series pertaining to crop production and many others. On the other hand department store sales expand before Easter and Christmas, a circumstance related to man-made festivals rather than to the weather. The production and sale of fireworks prior to Independence Day is of the same character. Sales of ice cream is an example of a partial shift from one type of seasonal to the other. In an earlier day sales were high in mid-summer and sank to a small fraction of the summer volume during the winter months. At that time ice cream was considered by most consumers as a luxury to be indulged in only when warm weather seemed to warrant its consumption. More recently ice cream has become a somewhat staple article of consumption. As a result the seasonal curve of sales has lost the greater part of its amplitude, albeit the peak is still in the summer months and the trough in the winter months.

The monthly sales of F. W. Woolworth Company for an eight-

year period are shown in Figure 75. This series exhibits seasonal variations accompanied by very little of either the trend or cyclical components. Two important features of the seasonal rhythm should be noted: (1) it recurs year after year with a fixed period; (2) the increases and decreases of sales occur at about the same time year after year. The seasonal rhythm has a fixed period and regular amplitude whereas the cyclical rhythm is variable in both characteristics.

Although the general seasonal pattern is clear enough in this series, there are discernible disturbances in some years. Particularly noticeable

FIGURE 75
MONTHLY SALES OF F. W. WOOLWORTH COMPANY 1930-37



Data from Table 115.

TABLE 115
SALES OF F. W. WOOLWORTH COMPANY, MONTHLY, 1930-37, INCLUSIVE*
(Millions of Dollars)

	1930	1931	1932	1933	1934	1935	1936	1937
January	18.4	19.2	18.0	15.8	18.1	17.1	17.0	18.7
February	20.0	19.4	18.8	16.3	17.9	18.2	19.0	19.8
March	22.5	21.7	21.3	17.5	24.0	20.5	19.7	24.8
April	24.4	23.8	20.8	20.2	19.8	22.4	23.1	21.9
May	25.3	24.1	20.5	19.8	22.0	21.1	22.6	24.6
June	20.7	22.0	18.9	19.3	22.0	21.1	23.4	24.2
July	20.7	21.1	18.1	19.6	19.5	20.2	22.9	24.7
August	22.1	21.7	18.2	20.4	20.8	21.6	23.2	22.8
September	22.4	21.7	19.5	21.6	21.3	20.2	23.4	24.3
October	26.4	26.2	22.5	22.0	23.3	23.4	26.7	26.8
November	24.1	22.0	20.2	21.0	22.3	23.4	23.9	25.1
December	42.3	39.7	33.1	37.0	39.6	39.6	45.5	47.2

* *Survey of Current Business* (Annual Supplements, 1932, 1936, and 1938), pp. 47, 26, and 28 respectively.

are the years 1933, 1934, and 1936. These were years of unsettled business conditions, hence the variations are to be expected. They are either cyclical or irregular in character and do not in any way affect the basic assumption that *seasonal variation is a component which repeats itself from year to year according to a regular pattern*. The method of determining the pattern for a particular series will be explained in a subsequent chapter.

Other Rhythms.—Some series contain other rhythmic movements which can be distinguished from the cyclical and seasonal rhythms because of differences in duration and cause. Chief of these are weekly rhythms and daily rhythms. The weekly rhythm is well illustrated by the flow of funds into and out of the Federal Reserve Bank of New York.

A regular though less spectacular currency movement is the typical weekly movement largely reflecting withdrawals for payrolls and the return of this currency as it is redeposited in banks. This movement is so regular that with a few exceptions one week is almost a repetition of another. On Thursday banks begin to withdraw from the Federal Reserve Bank the currency which their customers will require for their weekly payrolls and for week-end expenditures. Thursday and Friday there are large withdrawals for these purposes. Saturday there are smaller withdrawals. On Monday this money begins to flow back to the Reserve Banks and the flow continues on Tuesday and Wednesday.²

This weekly rhythm is shown in Figure 76 which is based on 26 weeks' experience at the Federal Reserve Bank of New York. "The period was one of increasing demands for currency and hence withdrawals somewhat exceeded deposits."³

Daily rhythms occur in such data as the hourly number of messages crossing a telephone switchboard, the hourly number of riders on trolley cars or the hourly use of electric power. These and many similar series have daily rhythms so regular that engineers use them to determine the amount of equipment to be kept in service each hour of the day and night. An illustration of the use of daily rhythms is afforded by the experience of a telephone company. A few years ago a very popular radio program was regularly broadcast from 7:00 to 7:15 in the evening. Up to 6:55 five generators were needed to carry the load of messages passing over the electric switchboard. By 6:58

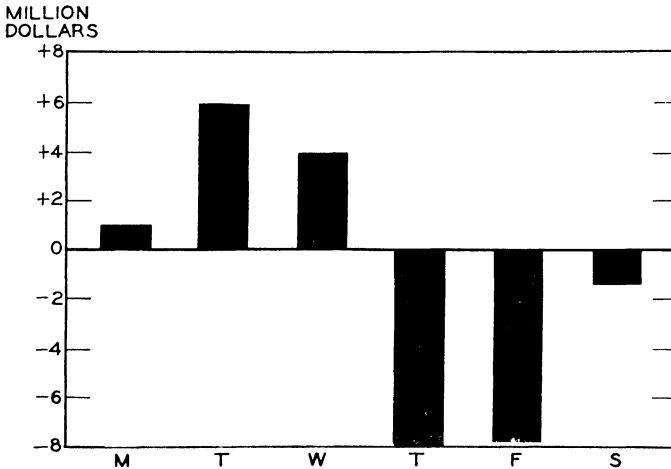
² W. Randolph Burgess, *The Reserve Banks and the Money Market* (New York: Harper & Bros., 1927), pp. 57-58.

³ *Ibid.*, p. 59.

FIGURE 76

DAILY NET CURRENCY MOVEMENT IN NEW YORK CITY TO OR FROM THE FEDERAL RESERVE BANK OF NEW YORK, APRIL TO SEPTEMBER, 1926

(+ indicates excess of deposits in the Reserve Bank and — excess of withdrawals from the Reserve Bank)



Data from W. Randolph Burgess, *Reserve Banks and the Money Market*, p. 60, with permission of the publishers, Harper & Bros.

the message rate would decrease to a point such that three generators were sufficient to carry the load. At 7:14 the other two generators had to be put back into use because the message load increased within five minutes after the end of the program to the highest point of the evening. From the advertising angle this story has interesting possibilities since it indicates indirectly the popularity of the radio program.

Many series of data contain no weekly or daily rhythms similar to those in the preceding examples. Even when shorter rhythms are present they have no common properties which hold from one series to another as is the case with seasonal and cyclical rhythms. The shorter rhythms do not require the use of statistical techniques beyond averages; consequently no further attention will be given to them

Irregular Factors

Determinate.—An irregular factor that can be dealt with statistically is called "determinate." Irregularity of the calendar is the only factor of this sort usually provided for in time-series analysis. The word "irregular" is used here to distinguish this movement in a time series from the trend component and the major rhythms—cyclical and seasonal

In using monthly data it is quite natural to think of the time unit as being constant. This is not the case, of course, since the number of days in a month varies from 28 to 31. But this is only one part of the variability. Some months have four Saturdays and Sundays, others have five. Some months have one or several legal holidays, others have none. Further, some series of data arise from activities which operate five days a week, others five and one-half or six or seven days. All of these factors tend to introduce variability into monthly data to an extent that is by no means negligible.⁴ The size of the corrections to be made for calendar irregularity and the method of making such corrections are discussed at the end of this chapter.

Indeterminate.—In this category are such things as weather disasters, labor strife, war, and all forms of unpredictable hazards. These things disturb the economic equilibrium and manifest themselves as irregular fluctuations in time series. The effect of three serious strikes in the anthracite coal industry can be seen in Part C of Figure 73. The declines in production during the years 1902, 1922, and 1925 are clearly not cyclical in character but are the direct result of enforced closing of the mines. The same sort of thing results from drought, flood, or earthquake. Wars also produce economic dislocations and lead to unexpected fluctuations in time series.

Disturbances of this kind do not always lead to declines such as those found in anthracite coal production. For example the drought in the cattle-raising states during the summer of 1934 resulted in a marked increase in the receipts of livestock at the Buffalo stockyards. Stock which ordinarily would have been marketed in the fall had to be moved off the land in the middle of the summer. Hence a weather disaster in the west led to an increase in receipts of livestock in the east.

In general the effects of unexpected events cannot be foreseen in advance of their occurrence any more than the events themselves can be foretold. In an analysis of a series into its components these movements are usually permitted to remain with the cyclical fluctuations. Therefore in what follows it will be understood that cyclical fluctuations will include whatever indeterminate irregular events may have occurred in any period for which a series is being studied.

⁴ Several plans have been devised in recent years to eliminate some of the calendar irregularity by an adjustment of the lengths of the several months or by the use of thirteen 28-day months. Some industrial concerns operate on a thirteen-period basis, but as a general rule the so-called "advantages" of calendar reform have been given scant attention by business men.

THE PROBLEM OF TIME SERIES ANALYSIS

The preceding part of this chapter consists of a statement of the components which may be present in a time series. Some series will contain all of the components, others only part of them. Some series are so largely controlled by one component that it is easily recognized from the original data, as was the case with the sales of the F. W. Woolworth Company in Figure 75. Usually the several components are not separately recognizable in the original data, but the business man wants to know the influence of each. Therefore, *the statistician's problem in dealing with time series is to develop methods of identifying the components and measuring them separately.*

The work of analysis can be divided into three parts: (1) preliminary analysis; (2) the measuring of trend; (3) the measuring of seasonal variation. In some of the methods which are discussed in succeeding chapters it is possible to reverse the order of steps 2 and 3, but step 1 always precedes the other two. After these three steps have been completed the cyclical fluctuations and the indeterminate irregular movements are the only ones remaining of the original list of components. These two stand as the residual of the statistical analysis. Sometimes it is possible to appraise in retrospect the effect of events that were currently indeterminate. But more commonly the two are interpreted together as the cyclical fluctuation of a series.

The remainder of this chapter and the three which follow contain an explanation of some of the methods available for carrying out all three steps in the analysis of series. In a particular business application only one or perhaps two of the steps may be needed. The various methods explained can be adapted readily to whatever form of partial analysis is required in a particular case.

PRELIMINARY ANALYSIS

The first step in the analysis of a time series is adjustment for the effect of calendar variation and change in the price level.⁵

Irregularity of the Calendar

The method of dealing with the variations in the number of days in the months of our calendar is to change each monthly total to an

⁵ It might be well for the student to understand that for some purposes it is desirable to allow the price cycle to remain with the residual cycle. As a general rule, however, a complete analysis of a series of data includes the adjustment for changes in the price level

average daily basis. The process itself is simple but in each case the proper number of days to be used must be counted in the calendar. The general rule is, count for each month the number of days that the activity was carried on during that month. In some cases this will mean all of the days in the months; in others Sundays, or Saturdays and Sundays will be excluded. Different holidays are observed in the several fields of business activity and in different communities. Consequently the number of days to be used will differ for the same series of data from one locality to another and in the same locality from one type of business to another.

Some examples will demonstrate the method of reducing to an average daily basis. In Part A of Table 116, column 1, the monthly sales of a Buffalo drugstore for 1936 are given. This store is open every day, hence the total number of days in each month has been used to change the sales to an average daily basis as shown in columns 2 and 3 of the table. In Part B of the table the amount of flour milled in Buffalo each month of 1936 is shown on an average daily basis. In this case Sundays and the six holidays, New Year's Day, Memorial Day, Independence Day, Labor Day, Thanksgiving, and Christmas, have been deducted from the number of days in the month to obtain the number of days flour mills were operated. In Part C of the table the same computation is shown for monthly bank clearings, but for this series, in addition to Sundays and the holidays listed in Part B, the following holidays have been deducted: Lincoln's Birthday, Washington's Birthday, Columbus Day, Election Day, and Armistice Day.

To show the effect of these corrections the monthly totals for each series and the daily averages are presented as index numbers in the last two columns of each part of the table. Columns 4, 9, and 14 have been obtained by dividing each monthly figure of columns 1, 6, and 11 by the simple average of the column. Columns 5, 10, and 15 have been obtained by dividing each daily average figure in columns 3, 8, and 13 by the weighted average of the column. The latter averages, \$218.82, 33,848 barrels, and \$5,585,000 have been computed by dividing the yearly total by the number of working days in the year for the respective enterprises.⁶

The effect of shifting to a daily average basis can be seen in Figure 77. In drugstore sales the changes due to the adjustment are very slight

⁶ This method is equivalent to weighting the daily average according to the number of working days in each month as explained under the rules for averaging ratios in chapter XI, p. 265.

TABLE 116
MONTHLY TOTAL AND DAILY AVERAGES FOR 1936 FOR THREE SETS OF DATA IN BUFFALO, NEW YORK:
SALES OF A DRUGSTORE, FLOUR MILLED, AND BANK CLEARINGS*

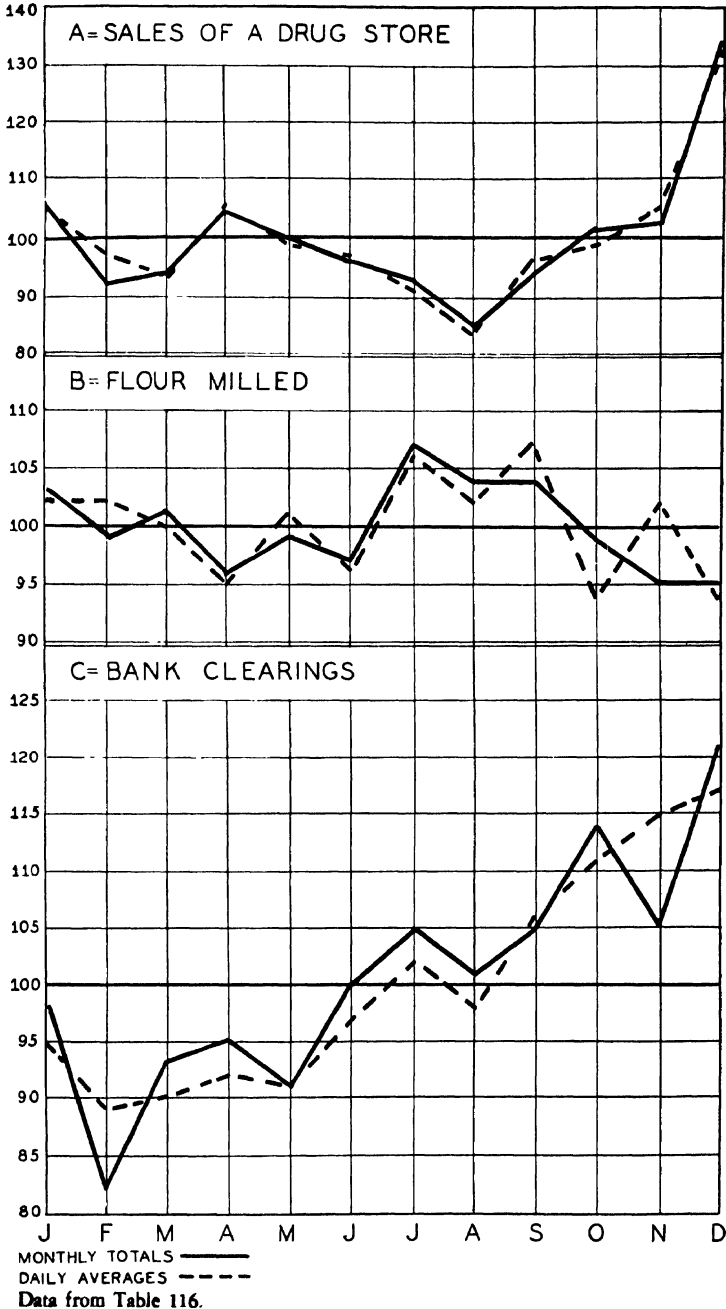
MONTH	A SALES OF A DRUGSTORE					B FLOUR MILLED					C BANK CLEARINGS				
	(1) Monthly Totals (dollars)	(2) of Work- ing Days in Month	(3) Average Sales Per Day (dollars)	(4) Index of Monthly Sales	(5) Index of Aver- age Daily Sales	(6) Monthly Totals (bbls.)	(7) No. of Work- ing Days in Month	(8) Average of Milled Per Day (bbls.)	(9) Index of Flour Milled per Month	(10) Index of Flour Milled Per Day	(11) Monthly Totals (million dollars)	(12) No. of Work- ing Days Per Month	(13) Average Clear- ings Per Day (million dollars)	(14) Index of Clear- ings Per Month	(15) Index of Clear- ings Per Day
January	7,094	31	228.84	106	105	897,115	26	34,504	103	102	138.6	26	5.33	98	95
February	6,137	29	211.62	92	97	862,208	25	34,488	99	102	115.1	23	5.00	82	89
March	6,303	31	203.32	94	93	880,138	26	33,851	101	100	131.4	26	5.05	93	90
April	6,918	30	230.60	104	105	833,387	26	32,053	96	95	134.0	26	5.15	95	92
May	6,682	31	215.55	100	99	858,749	25	34,350	99	101	127.8	25	5.11	91	91
June	6,378	30	212.60	96	97	843,536	26	32,444	97	96	141.7	26	5.45	100	97
July	6,193	31	199.77	93	91	931,902	26	35,842	107	106	148.3	26	5.70	105	102
August	5,647	31	182.16	85	83	900,662	26	34,641	104	102	142.1	26	5.47	101	98
September	6,275	30	209.17	94	96	905,020	25	36,201	104	107	141.1	25	5.64	100	101
October	6,714	31	216.58	101	99	860,082	27	31,855	99	94	161.1	26	6.20	114	111
November	6,794	30	226.47	102	104	825,323	24	34,388	95	102	141.1	22	6.41	100	115
December	8,952	31	288.77	134	132	927,213	26	31,816	95	94	170.0	26	6.54	121	117
Average	6,674	...	218.82	...	...	868,778	...	33,848	...	...	141.0	...	5.59	...	...

* From the files of the Bureau of Business and Social Research of the University of Buffalo

FIGURE 77

MONTHLY TOTALS AND DAILY AVERAGES FOR 1936 FOR THREE SETS OF DATA IN
BUFFALO, NEW YORK: SALES OF A DRUG STORE, FLOUR MILLED AND
BANK CLEARINGS

INDEX



except in February. The adjustment of flour milling is more important because of the greater variation between the number of days in a month and the number of days flour mills were operated during that month. The months of September, October, and November are particularly noteworthy because the monthly totals conceal irregularities that are brought to light by the daily averages. A reverse tendency appears in bank clearings, the daily average curve eliminating some of the artificial irregularity introduced by the monthly total curve.

Other examples could be added with calendar corrections different from the three presented here, but these are typical of the corrections that will be needed in practice. These three examples in the order presented are sometimes referred to as, (1) correction for all days, (2) correction for working days, (3) correction for banking days. This distinction seems unnecessary if the general rule is followed, namely, use the number of days in each month during which the activity was in operation. This rule can usually be applied without difficulty in spite of the effect upon operating schedules of the Wage and Hour Law and the requirements of the defense program.

The method of reducing to an average daily basis should be used when the monthly total is made up of non-overlapping daily items. In the examples used, the drugstore sales, flour milled, and value of checks cleared, respectively, were distinct for each day's business. There was no duplication from day to day. On the other hand series such as the number of employees of drugstores, the capacity of flour mills, or the deposits of banks should not be reduced to an average daily basis, because the same employees are in the drugstores, the same machinery is in the flour mills, and the same deposits are in the banks for all or at least part of the month. Such data should be recorded as of a specified date each month.

Finally some series are collected on a weekly basis. Usually such series are put on a monthly basis by taking the average of either four or five weeks as the calendar may happen to work out.⁷ This weekly average is a satisfactory substitute for an average daily basis under ordinary circumstances. A difficulty arises when working time during a week is curtailed by a holiday. The proper way to deal with this situation is to adjust the week in question to an equivalent full-time basis.

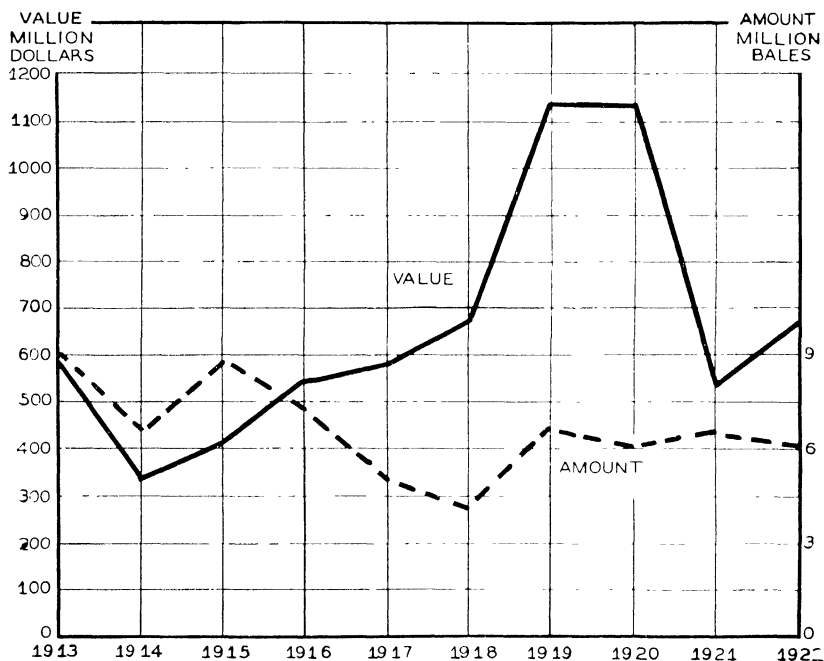
⁷ The method of doing this is illustrated for bank debits in chapter XX, Table 112

Change in the Price Level

The effect of changing prices can be seen in Figure 78 which shows United States exports of cotton in bales and in dollar value. The value of cotton exported increased tremendously during the war and in the years immediately following. These figures are often referred to as evidence of the expansion of our export trade in cotton. Yet the facts are that the amount of cotton exported declined during the war years and in no subsequent year was it as high as in 1915. The value increased with the increase in price, but the amount of cotton exported decreased.

Since both value and quantity are available in this case, it can be used as a controlled experiment to demonstrate the method of correcting for price change. The export price of cotton should be used in making the correction. In Table 117, the value of cotton exported in column 1 is divided by the average export price in column 2 to obtain the quantity exported in column 3. The quantity exported in pounds

FIGURE 78
RAW COTTON EXPORTED BY THE UNITED STATES, 1913-22



Data from Table 117.

TABLE 117
RAW COTTON EXPORTED BY THE UNITED STATES, 1913-22

YEAR	(1) VALUE OF COTTON EXPORTED * (000 omitted)	(2) AVERAGE EXPORT PRICE (cents per pound) †	(3) COMPUTED QUANTITY EXPORTED (million pounds) (1) ÷ (2)	(4) COMPUTED QUANTITY EXPORTED (thousand bales)	(5) ACTUAL QUANTITY EXPORTED ‡ (thousand bales)
1913	\$575,496	12.4	4,641	9,282	8,964
1914	343,908	10.7	3,214	6,428	6,565
1915	417,012	10.4	4,010	8,020	8,724
1916	545,232	15.2	3,587	7,174	7,291
1917	575,304	23.8	2,417	4,834	4,952
1918	674,124	30.2	2,232	4,464	4,235
1919	1,137,372	33.9	3,355	6,710	6,735
1920	1,136,412	36.0	3,157	6,314	6,159
1921	534,240	16.2	3,298	6,596	6,474
1922	673,248	21.6	3,117	6,234	6,114

* *Survey of Current Business* (Annual Supplement, 1936), p. 67.

† *Statistical Abstract* (1921), p. 621 and (1936), p. 309. (Prices for fiscal years averaged to obtain prices for calendar years.)

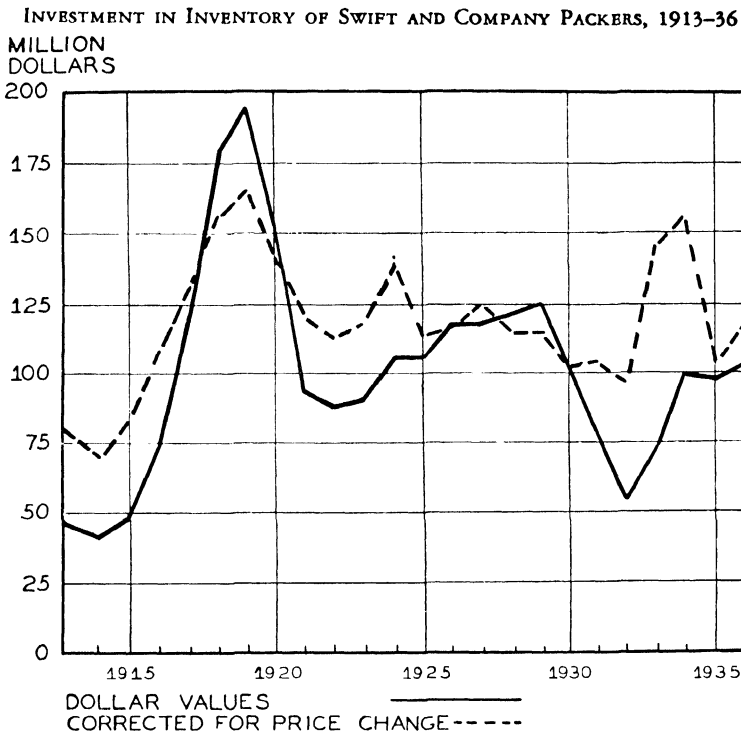
‡ "Record Book of Business Statistics," *Survey of Current Business* (1927), p. 26.

is then divided by 500 to reduce it to equivalent bales as shown in column 4. Column 5 gives the number of bales exported as reported directly by the Department of Commerce, and therefore serves as a check on the validity of the results in column 4. Comparison of the two columns shows how the method of correcting for price change has worked out. The greatest differences appear in 1913 and 1915, but in none of the years is there perfect agreement. Three reasons can be advanced in explanation of the differences: (1) total value of exports for a year divided by an average of monthly prices would give the exact quantity exported only in case an equal amount were exported each month; (2) variations in the quality of cotton exported in different months would tend to spread prices and accent the discrepancies arising from variation in quantity exported from month to month; (3) prior to 1916, export prices are not quoted satisfactorily in the source books.

When the series expresses the dollar value of a number of different products, the correction for price change cannot be made by the use of individual prices as was done in the preceding example. Instead a price index must be used as is demonstrated in the following example. The value of the inventory of meat products of Swift & Company at the close of each year is given in Table 118. The question is, to what extent are the variations from year-to-year due to changes in the prices at which the inventory is valued and to what extent are they due to a greater or smaller amount of meat stored? The correction for price

change is made by dividing the inventory by the index of prices of meat products taken from the Bureau of Labor Statistics Index of Wholesale Prices. The effect of the correction is shown in Figure 79. A considerable part of the change in value of inventory during the war years was due to price, although the inventory increased in amount to the extent shown by the dotted line on the chart. After 1921 the corrected inventory tended to follow a regular course except in the years 1924, 1933, and 1934.

FIGURE 79



Data from Table 118.

Conclusion

The reduction of data to an average daily basis and the correction for price change, if the series represents dollar values, eliminate the effect of two of the components which may be present in a time series. The data are then ready for separation into the trend, seasonal and cyclical components. The methods of doing this will be presented in the next three chapters.

TABLE 118
INVESTMENT IN INVENTORY OF SWIFT AND COMPANY PACKERS
AT THE CLOSE OF EACH YEAR, 1913-36

YEAR	(1) INVENTORY * (millions)	(2) INDEX OF WHOLESALE PRICES OF MEAT PRODUCTS † (1926 = 100)	(3) CORRECTED INVENTORY (1) ÷ (2) (millions)	YEAR	(1) INVENTORY * (millions)	(2) INDEX OF WHOLESALE PRICES OF MEAT PRODUCTS † (1926 = 100)	(3) CORRECTED INVENTORY (1) ÷ (2) (millions)
1913...	\$47.0	59.8	\$78.6	1925...	\$105.0	93.3	\$112.5
1914...	44.0	62.6	70.3	1926...	115.5	100.0	115.5
1915...	48.0	57.6	83.3	1927...	116.0	92.7	125.1
1916...	75.0	66.4	113.0	1928...	121.5	107.0	113.6
1917...	120.0	92.9	129.2	1929..	124.5	109.1	114.1
1918...	178.0	115.2	154.5	1930...	100.0	98.4	101.6
1919...	192.5	117.6	163.7	1931...	77.5	75.4	102.8
1920...	152.0	108.0	140.7	1932..	56.0	58.2	96.2
1921...	93.5	77.4	120.8	1933..	72.0	50.0	144.0
1922...	86.0	76.6	112.3	1934..	99.5	62.9	158.2
1923...	89.5	76.2	117.5	1935..	96.5	94.5	102.1
1924...	104.5	75.7	138.0	1936..	102.5	87.8	116.7

* *Swift & Company Year Book* (1936).
† *Survey of Current Business* (1936 Supplement).

PROBLEMS

1. Why do most series of data representing fundamental factors of our economic system have positive trends?
2. Does long-time trend have a constant influence on a series of data? Explain.
3. The annual production of electricity in the United States from 1920 to 1940 was as follows (billion k.w.h.):

YEAR	PRODUCTION OF ELECTRICITY	YEAR	PRODUCTION OF ELECTRICITY
1920	43	1931	91
1921	41	1932	82
1922	48	1933	85
1923	56	1934	91
1924	59	1935	98
1925	66	1936	112
1926	74	1937	122
1927	79	1938	117
1928	87	1939	130
1929	96	1940	145
1930	95		

- a) Plot these data.
 - b) Indicate the trend of electricity production and justify your trend.
4. Take from published sources three series of data extending over at least ten years. Indicate the location of the trend of each series explaining your choice.
 5. How are cyclical fluctuations and seasonal variations distinguished?

6. What is meant by the amplitude of the cycle? the period? phase?
7. Which of the following series would have seasonal variations related to natural forces? (a) sales of rubber footwear; (b) production of cement; (c) income of hotels in Florida; (d) sales of automobiles in California; (e) postal receipts.
8. Which of the series in Problem 7 would have cyclical movements resulting directly from changes in the price level?
9. What is the effect of the Wage and Hour Law upon the adjustment of a production series for calendar variation? The effect of three-shift, seven-day operation during the period of defense preparation?
10. List examples (other than those in the text) of series affected by indeterminate irregularities.
11. What is the central problem of time-series analysis?
12. Could the same number of working days per month be used in reducing to an average daily basis the monthly production of cement in Pennsylvania and the monthly production of rubber tires in Ohio? Why or why not?
13. Which of the following should be changed to an average daily basis and which should not? Explain in each case.
 - a) Average monthly sales per sales person in a department store.
 - b) A monthly record of the stocks of finished cement in warehouses.
 - c) The number of employees on the payroll of an industrial concern on the 15th of each month.
14. Using Figure 79 as a basis, write a statement of the importance of price change and the importance of other factors as causes of cyclical fluctuations.

REFERENCES

- CROXTON, FREDERICK E., and COWDEN, DUDLEY J., *Applied General Statistics*. New York: Prentice-Hall, Inc., 1939. Chapter XIV.
- MITCHELL, WESLEY C., *Business Cycles, The Problem and Its Setting*. New York: National Bureau of Economic Research, 1927.
- The basic problem of analyzing time series is discussed on pages 189–212

CHAPTER XXII

TREND

ANYONE can grasp the notion that there are fundamental forces present in our economic system which manifest themselves over a period of years in the form of growth or decline in series of basic data. The tendency to growth or decline spreads to interrelated series in every direction so that the trend component of series becomes characteristic of the system. There is, however, no such thing as a rate of change applicable to all series for a given period of time. The trend of each series must be determined separately.

Recognition of the existence of the trend component and recognition of the desirability of measuring it are merely the preliminaries to the statistician's two major problems in dealing with the subject. The two problems are the determination of the path followed by the trend of a given series and the selection of a method of measurement which will provide numerical values along this path throughout the period of the series. These two problems are usually known as the location of trend and the measurement of trend

THE LOCATION OF TREND

At this stage of the analysis of a series the effects of calendar variation and price change have been adjusted. The remaining components are trend, seasonal variation (except in annual data), cyclical fluctuations, and residual irregularities.

The most effective method of locating a trend is by constructing a graph of the data. Such a graph may employ an ordinary, a logarithmic, or other specialized scale. In every case the object is to reproduce the cyclical component of the data in a form which will facilitate the location of the trend. The actual process consists in studying the graph to establish the positive and negative phases of the cycles of the given series. This must be done in the light of the observer's knowledge of the particular series and of general business conditions, and when located on the graph is known as a free-hand trend.¹

¹ Sometimes no measurement of trend is undertaken, reliance being placed solely on the subjective free-hand determination. Such procedure has usually proved unsatisfactory for long periods but can sometimes be used for short periods, as will be apparent later in the chapter.

Ordinarily judgments of the location of the trend of completed cycles cause no great difficulty because the surrounding facts can be ascertained in case of doubt. The same certainty does not exist with reference to current data included in a series. Specifically the trend of a series running from 1920 to date (1941) could be located readily as far as 1930 or 1931, but the path to be selected for subsequent years must remain in some doubt until the cyclical movement of the 1930's is completed.

The process of locating the trend of a series and measuring trend values to be used in analyzing the series is usually referred to as fitting a trend. This is a defensible but somewhat extended use of the word "fit." The object is not to obtain a trend which will "fit" the data in the sense of falling as close as possible to all of the original values. The extreme of this, of course, would be a reproduction of the original data. Further, trend does not mean a curve which, while smoothing out minor irregularities of the data, follows or fits the main contours of those data. On the contrary the trend component is described as an average position with respect to the rhythmic movements of the data. It may be measured so as to be affected very little by the wave movements, or some flexibility² of direction may be introduced in order to reproduce in the trend component traces of the direction of the wave movements. The question of fit and the matter of flexibility will reappear on subsequent pages in the evaluation of the methods of measuring trend.

The location of trend is a subjective process inasmuch as the observer uses his individual judgment aided by the best collateral knowledge available to determine that the trend of a series shall follow a chosen path. The second step, measuring the trend, consists of implementing the subjective judgment of the position of the trend by an objective determination of its successive values during any given period of time. The extent to which the various methods of trend measurement blend the objective-subjective elements will be discussed along with the presentation of the details of those methods.

METHODS OF MEASURING TREND

Once the path of the trend of a series has been located and the decision has been made to establish that path by an objective measure

² A flexible measure of trend is one that can change direction, such as a moving average. An inflexible measure of trend is one which cannot change direction, such as a straight line, a compound interest curve, or a power curve.

TABLE 119
COMPUTATION OF NINE-YEAR MOVING AVERAGE FITTED TO CONTROLLED DATA
CONTAINING CYCLE AND TREND—28-YEAR PERIOD

Year	(1) Orig- inal Cycle	(2) 9-Year Moving Total	(3) 9-Year Moving Average	(4) Straight- line Trend	(5) Trend and Cycle (1) + (4)	(6) 9-Year Moving Total	(7) 9-Year Moving Average	(8) Trend and Cycle (1) × (4)	(9) 9-Year Moving Total	(10) 9-Year Moving Average	(11) Curved Trend 5 Per Cent Annual Increase	(12) Trend and Cycle (1) × (11)	(13) 9-Year Moving Total	(14) 9-Year Moving Average
1st	100			100	200			100.0			100	100.0		
2d	104			105	209			109.2			105	109.2		
3d	110			110	220			121.0			110	121.0		
4th	107			115	222			123.0			116	124.1		
5th	102	900	100	120	222	1,980	220	122.4	1,074.9	119.4	122	124.4	1,097.7	122.0
6th	98	900	100	125	223	2,025	225	122.5	1,119.9	124.4	128	125.4	1,152.7	128.1
7th	93	900	100	130	223	2,070	230	120.9	1,166.7	129.6	134	124.6	1,213.0	134.8
8th	90	900	100	135	225	2,115	235	121.5	1,216.2	135.1	141	126.9	1,280.1	142.2
9th	96	900	100	140	236	2,160	240	134.4	1,264.4	140.5	148	142.1	1,348.6	149.8
10th	100	900	100	145	245	2,205	245	145.0	1,310.3	145.6	155	155.0	1,417.0	157.4
11th	104	900	100	150	254	2,250	250	156.0	1,354.4	150.5	163	169.5	1,485.6	165.1
12th	110	900	100	155	265	2,295	255	170.5	1,396.3	155.1	171	188.1	1,554.4	172.7
13th	107	900	100	160	267	2,340	260	171.2	1,436.8	159.6	180	192.6	1,623.7	180.4
14th	102	900	100	165	267	2,385	265	168.3	1,480.0	164.4	189	192.8	1,701.4	189.0
15th	98	900	100	170	268	2,430	270	166.6	1,525.0	169.4	198	194.0	1,787.4	198.6
16th	93	900	100	175	268	2,475	275	162.8	1,571.8	174.6	208	193.4	1,881.0	209.0
17th	90	900	100	180	270	2,520	280	162.0	1,621.3	180.1	218	196.2	1,984.4	220.5
18th	96	900	100	185	281	2,565	285	177.6	1,669.5	185.5	229	219.8	2,090.3	232.3
19th	100	900	100	190	290	2,610	290	190.0	1,715.4	190.6	241	241.0	2,196.4	244.0
20th	104	900	100	195	299	2,655	295	202.8	1,759.5	195.5	253	263.1	2,303.3	255.9
21st	110	900	100	200	310	2,700	300	220.0	1,801.3	200.1	265	291.5	2,409.4	267.7
22d	107	900	100	205	312	2,745	305	219.4	1,841.8	204.6	279	298.5	2,518.3	279.8
23d	102	900	100	210	312	2,790	310	214.2	1,885.0	209.4	293	298.9	2,640.3	293.4
24th	98	900	100	215	313	2,835	315	210.7	1,930.0	214.4	307	300.9	2,772.3	308.0
25th	93			220	313			204.6			322	299.5		
26th	90			225	315			202.5			339	305.1		
27th	96			230	326			220.8			356	341.8		
28th	100			235	335			235.0			373	373.0		

rather than to rely upon a free-hand determination, the next problem is to find what measure will accomplish the purpose. The choice can be made from a variety of devices which have been developed in the past. Many of these, such as the joining of selected points, the joining of a series of averages, or taking the averages of the high and low points of cycles, are not in sufficiently general use to warrant explanation here. A large part of actual practice will be covered by explaining two types of trend measurement, the moving average and curves based on elementary mathematical functions.

Moving-Average Trend

The moving average is computed in the same way as any other average. The measurement of trend results from putting together a series of simple averages, hence the name moving average. For example a five-year moving average means that the average of the data for the first five years is taken as the trend value for the third year, that the average of the data for the second to sixth years is taken as the trend value for the fourth year, etc. The principle on which this computation is based is the rhythmic movement of cycles. If a series of data contains cycles of five years' duration, any five-year moving average computed from the series will therefore contain one complete cycle, and the average will be free of cyclical influence since the fluctuations in the two directions offset each other.

The properties of the moving average can be explained with the aid of a set of data constructed for the purpose. Column 1 of Table 119 contains a series of regular nine-year cycles written above and below 100 as a standard or normal. Column 2 gives the moving nine-year total and column 3 the nine-year moving average. This computation shows that the moving average removes the cyclical rhythm which was put into the data originally.

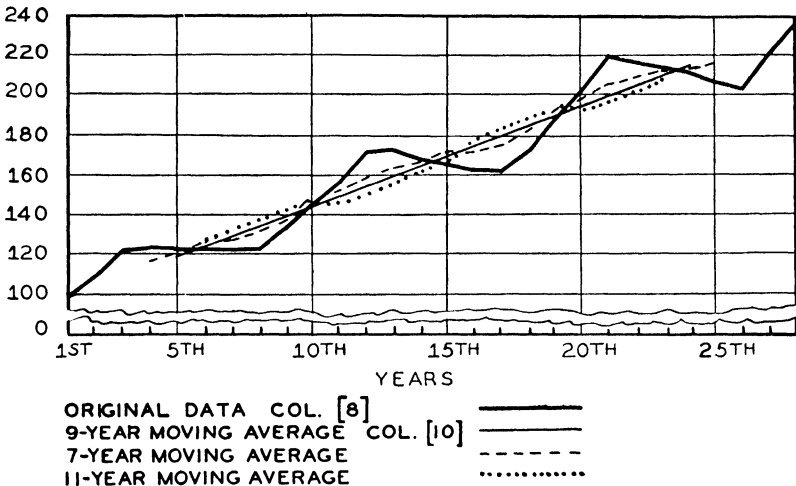
A question may now be raised as to how the moving average will behave when trend is present in the data. That is, if a known trend and a known cyclical fluctuation are combined, will a properly selected moving average computed for the combined series separate the two components which it contains? This question is answered in the second part of Table 119. Column 4 gives a straight-line trend with a yearly increment of five. In column 5 this trend and the cycle from column 1 have been added to obtain a series combining the two components. A nine-year moving average has been fitted to this series as shown in

columns 6 and 7. The moving average values are exactly the sum of the trend and the normal value of the cycle. Hence the moving average has separated the two components.

The trend and cyclical components can also be combined by multiplication as shown in column 8 of Table 119 for the controlled series. Specifically each item of column 1 is multiplied by the corresponding item of column 4, the former being considered as per cents. The cycles of the combined series increase in amplitude as can be seen in Figure 80. A nine-year moving average has been fitted to the series in col-

FIGURE 80

MOVING AVERAGES FITTED TO CONTROLLED DATA CONTAINING CYCLE AND STRAIGHT LINE TREND

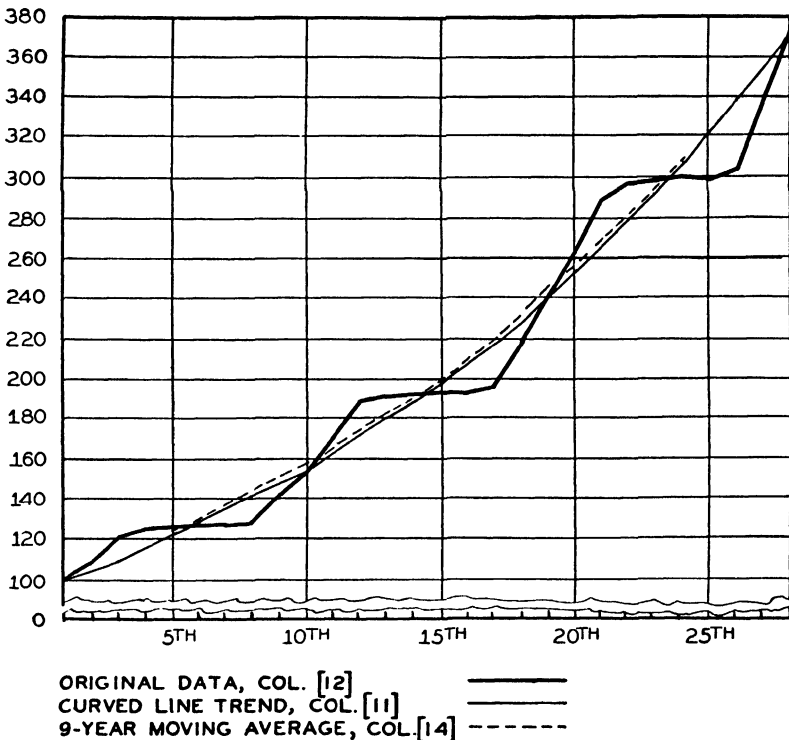


umns 9 and 10. The trend obtained in column 10 is not an exact reproduction of column 4. This difference is not merely a matter of rounding-off numbers but is the result of changing to relative cycles. When the cycle is changing from 90 up to 110 the moving-average values are slightly above the assumed trend. When the cycle is changing from 110 down to 90 the moving-average values are slightly below the assumed trend. The variation between the computed and the assumed trend is not to be taken as a serious defect of the moving average. In this case the variations are negligible and even though the cycles had much greater amplitude than those used here, the error involved would be less important than certain other factors to be considered in succeeding pages.

The next question for discussion is, How does the moving average behave when the assumed trend is not a straight line? To test this a compound-interest curve with an annual rate of growth of 5 per cent is introduced in column 11 of Table 119. Column 12 gives the results of multiplying each of these trend values by the values of column 1 considered as relative cycles. That is, column 12 gives a new series combining trend and cycles. A nine-year moving average has been fitted to the series in columns 13 and 14. The results are shown in Figure 81. The computed trend is consistently higher than that as-

FIGURE 81

MOVING AVERAGE FITTED TO CONTROLLED DATA CONTAINING CYCLE AND CURVED LINE TREND



Data from Table 119.

sumed in the series. This situation will always arise when the series contains a growth component which is increasing by more than constant amounts from year to year. If the series has a trend which is decreasing by more than constant amounts the moving average will

give a result which is consistently below the true values. The error introduced, while more serious than in the preceding case, is again not important enough to warrant the abandonment of moving-average trend. Some authors have gone so far as to say that the moving average should not be used unless the true trend is approximately a straight line. Such a limitation placed upon the use of the moving average seems unwarranted in view of the slight error introduced when it is used with a series such as that shown in Figure 81. Few cases will be found in practice with growth component exceeding the 5 per cent rate used in making the data plotted on this chart.

Selection of Length of Period.—The period of the moving average is the number of years included in computing each average. Thus a nine-year period is used in Table 119. In dealing with actual data the number of years to include in the average will not be known, as it was in the controlled example, but must always be determined from study of the data. If the number taken is less than the length of the cyclical fluctuations, part of the cycle will remain in the moving average. If the number of years taken exceeds the length of the cycle, then inverse movement will be introduced into the moving average. These statements can be verified by reverting once more to the controlled data. Seven-year and eleven-year moving averages⁸ computed from column 8 of Table 119 are shown in Figure 80. The seven-year curve bends mildly with the cyclical fluctuations. We know that the trend should be a straight line; therefore the use of this seven-year curve as the trend would mean that part of the cycle was being measured as trend. The eleven-year curve contains wave movements which are inverse to the direction of the cyclical fluctuations. If it were taken as the trend, the amplitudes of the cycles would be magnified because of the inverted path followed by the trend.

The situation found here is not peculiar to the controlled data but applies to every case of fitting a moving average. Hence the rule: *the period of a moving average used to measure trend should be equal to the length of the cycles present in the data or some integral multiple thereof.* If the data contain five-year cycles, trend can be measured by a five-year, a ten-year, etc., moving average, but by no moving average with period other than a multiple of five.

Cycles of different period: The use of the foregoing rule would be straightforward, if the cycles of a series of data were all of the same

⁸ The computations have been omitted since the method is now familiar to the reader.

length. It was shown in the preceding chapter that this is not the case. Hence the question arises, what is to be taken as the length of the period of a moving average fitted to a series containing cycles which may vary in period from three years to eight or ten years? The only way to deal with this situation is to take as the period the number of years which best conforms to the cycles in the data. The lengths of the various cycles should be tabulated from peak to peak and from trough to trough. The modal length of cycle appearing in the tabulation should be taken as the period. This means, of course, that those cycles having less than the modal period will be over-corrected, while those having greater than the modal period will be under-corrected. Hence the moving average is only an approximate measure of trend.

When the period contains an even number of years: The examples of moving average computation that have been given up to this point have all had an odd number of years in the period. When an even number of years is used, an additional step must be introduced. This is known as centering the moving average. The process can be explained best with the aid of an example.

TABLE 120

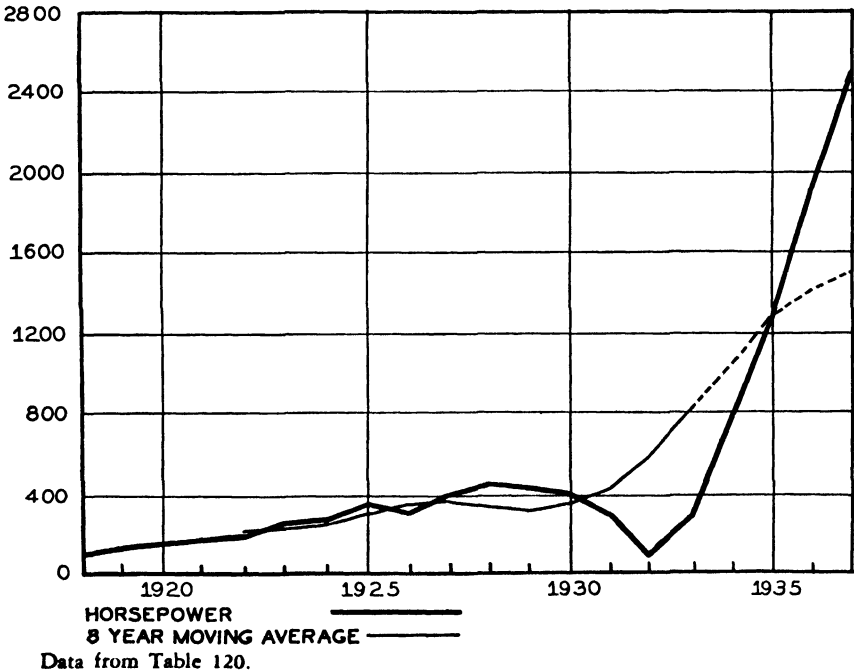
NUMBER OF HORSEPOWER OF DIESEL ENGINES INSTALLED ANNUALLY IN THE UNITED STATES, 1918-37* WITH EIGHT-YEAR MOVING-AVERAGE TREND AND FREE-HAND PROJECTION OF THE TREND THROUGH THE YEARS 1934-37

YEAR	(1) NEW DIESEL HORSE- POWER INSTALLED (000 omitted)	(2) 8-YEAR MOVING TOTAL	(3) 8-YEAR MOVING AVERAGE	(4) 2-YEAR MOVING TOTAL OF 8-YEAR MOVING AVERAGES	(5) 8-YEAR CENTERED MOVING AVERAGE
1918.....	100				
1919.....	125				
1920.....	137				
1921.....	140				
1922.....	185	1,567	196	418	209
1923.....	250	1,772	222	478	239
1924.....	280	2,051	256	552	276
1925.....	350	2,364	296	628	314
1926.....	305	2,654	332	691	346
1927.....	404	2,873	359	725	362
1928.....	450	2,928	366	709	354
1929.....	430	2,746	343	677	338
1930.....	404	2,676	334	677	338
1931.....	305	3,121	390	724	362
1932.....	98	3,967	496	886	443
1933.....	280	5,417	677	1,173	586
1934.....	750	7,487	936	1,613	806
1935.....	1,250				1,070
1936.....	1,900				1,260
1937.....	2,500				1,400
					1,490

* *The Annalist*, Vol. 51, No. 1325 (June 10, 1938), p. 787.

Table 120 contains the figures for horsepower of Diesel engines sold each year for 20 years. The outstanding feature of this series is the tremendous increase since 1933, but other movements occur in the data, as may be seen in Figure 82. A moving average with an eight-year period has been fitted to the data to secure a smooth trend. The moving eight-year total and average are shown in columns 2 and 3. These figures have been set down midway between the positions of the original data to indicate that the center of an eight-year period falls midway between the fourth and fifth years. For graphic presentation of the relation between data and trend the intermediate position of the trend values with respect to the values of the data makes no difference. When the trend component is to be removed from the data, trend values are needed which coincide in time with the values of the original data. The intermediate trend values are centered on the original data by simple straight-line interpolation performed by summing the eight-year averages in pairs as shown in column 4 of the table and dividing each figure of column 4 by two to obtain the eight-year centered moving-average trend of column 5.

FIGURE 82
NUMBER OF HORSEPOWER OF DIESEL ENGINES INSTALLED ANNUALLY, 1918-37
THOUSANDS



The original data and the trend are plotted in Figure 82. The trend is a satisfactory fit except in 1932 and 1933 when it appears to be affected too much by the high figures of 1935, 1936, and 1937. It is, of course, quite possible that new installations of Diesel horsepower in future years will continue at the 1937 level or higher. Even so, the trend turns up too sharply in 1932 and 1933 to be a satisfactory description. The absence of trend values for the first four and the last four years is also objectionable, but this is inevitable in the moving-average method. A free-hand trend for the missing years has been indicated on the chart.

Considerable time can be saved in the computation of the moving average by the use of a listing-adding machine. The computation of column 2 is shown on the adding machine record reproduced at the right. At each step after the first total (subtotal on the machine) one figure is subtracted and one added to obtain the new eight-year total (subtotal on the machine). At the end, the last eight figures should be added independently as a check.

Keeping the Moving Average Current.—The moving average has a great disadvantage for use in current computations: the trend cannot be computed for one or more recent years depending on the length of the period. Several methods of supplying the missing trend figures by weighting the data of the last few years have been proposed but they have not proved satisfactory. The best method of keeping a moving-average trend up to date is to extend it free hand. The exercise of ordinary care by one familiar with the data may lead to a free-hand trend which is good enough for temporary use. The estimate should be adjusted as soon as enough data become available. The choice here is between a free-hand method which although solely dependent on an observer's judgment is likely to be nearly correct, and a mathematical method which *may* or *may not* give a better result than the free-hand method, but has the possibility of being entirely misleading. Of the two the free-hand method seems to be less dangerous, hence its use is preferred. The last four entries in column 5 of Table 120 are the values of a free-hand trend read from Figure 82. It would be a valuable exercise for the reader to collect the data for subsequent years and

100
125
137
140
185
250
280
350
1,567 s
100 —
350
1,772 s
125 —
404
2,051 s
137 —
450
2,364 s
140 —
430
2,654 s
185 —
404
2,873 s
250 —
305
2,928 s
280 —
98
2,746 s
350 —
280
2,676 s
305 —
750
3,121 s
404 —
1,250
3,967 s
450 —
1,900
5,417 s
430 —
2,500
7,487 *
404
305
98
280
750
1,250
1,900
2,500
7,487 *

bring the moving average up to date in order to determine how much error was involved in using these free-hand values.

Advantages of the Moving Average.—(1) The principle of the moving average as a measure of trend is easy to understand. (2) The computation, although somewhat laborious, is simple in character. (3) The greatest advantage is flexibility. When some major change in the growth component of a series occurs, the moving average automatically accommodates itself to the change. This feature is particularly valuable in fitting trends to the years since 1929. Undoubtedly many series will exhibit growth components in subsequent years which will have little relation to those shown by the same series before the depression. It will appear later in the chapter that mathematical curves may not be sufficiently flexible to accommodate themselves to this situation. On the other hand the moving average will give an adequate description of such changes. (4) Trend measured by a moving average can never leave the data. That is, the trend is measured as an average situation and the average must give a value between the limits of the data from which it is computed.

Defects of the Moving Average.—(1) The values of the trend at the beginning and end of the series cannot be computed. The loss of data at the beginning of the series is usually not very serious, but the loss of data for the most recent years is the greatest defect of the moving average. For this reason the method of supplying the trend for the missing years free hand has been proposed. (2) If the moving average is to give a correct measure of trend the cycles must have a uniform period. But this is never true in practice; therefore the moving average has variable accuracy as a measure of trend. The greater the difference between the period of a given cycle and the period of the moving average, the less accurate the representation of trend. In particular the long cycle which we are experiencing during the decade 1930–40 requires a longer period moving average than was customary in previous years. (3) The moving average will not represent accurately a trend which increases or decreases constantly according to some law other than a straight line, as was shown in columns 11 to 14 of Table 119 and in Figure 81. This defect is of greater interest from the theoretical than the practical point of view. The error introduced is decidedly less important than that arising from variation in the period of cycles, which may cause appreciable over- or under-correction of the cyclical fluctuations.

Curves Based on Mathematical Functions

If the growth of a series follows a fixed law, that law can be expressed by a mathematical function. This is the basis on which statisticians have spent considerable energy in adapting various types of mathematical curves to the measurement of trend in time series. This work has been extremely fruitful in past years, although somewhat less so since 1929. At one time after World War I it appeared that the trends of most general series could be represented by straight lines. Series descriptive of operations of individual concerns sometimes required more complicated curves, but for the most part trend was expressed by a straight-line equation. After 1929 the disadvantage of complete reliance upon either straight lines or more complex curves became apparent. At the present time the general consensus would be that mathematical curves can be used to measure trend in some cases but they lack the flexibility necessary for proper description of current trend movements. The reader should not assume from the foregoing statement that mathematical measurement of trend can be discarded. On the contrary the concept of measurement according to a fixed law is important in understanding the nature of trend, even though the current practice is to consider the trend component as variable both in intensity and direction. Accordingly a discussion of the more commonly used mathematical curves is in order.

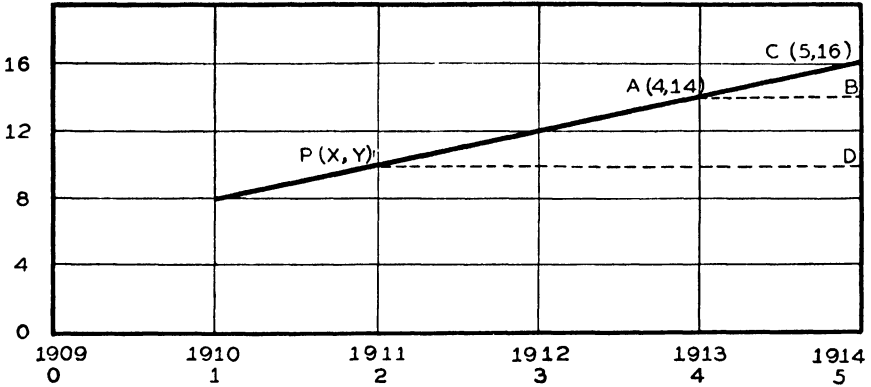
Straight Line.—Suppose that we have the following data:

YEAR	SALES
1910	8
1911	10
1912	12
1913	14
1914	16

The data are plotted in Figure 83 and, of course, they lie on a straight line. In plotting these points two axes of reference were used to locate the points. The time intervals (abscissas) were placed from left to right on the horizontal or X -axis and the sales (ordinates) were located according to the vertical scale or Y -axis. This terminology will be used to develop the method of writing an equation for the line on which these points lie. If 1909 is taken as the origin on the X -axis, the abscissas of the five years will be 1, 2, 3, 4, and 5 as indicated on the chart. Now take any point P on the line. Its abscissa will be X and its ordinate Y . The three points $P (X,Y)$, $A (4,14)$, $C (5,16)$

FIGURE 83

DIAGRAM USED TO WRITE THE EQUATION OF A STRAIGHT LINE



can be used to write the equation of the line. The triangles ABC and PDC are similar; therefore $\frac{BC}{AB} = \frac{DC}{PD}$. Substituting the values of the points in this equation gives

$$\frac{16 - 14}{5 - 4} = \frac{16 - Y}{5 - X} \text{ or } \frac{2}{1}(5 - X) = 16 - Y;$$

hence $Y = 6 + 2X$. The value of Y on the line corresponding to any value of X can be obtained by substituting the value of X in the equation. Thus

$$\begin{aligned} \text{when } X &= 0, & Y &= 6 \\ \text{when } X &= 1, & Y &= 6 + (2 \times 1) = 8 \\ \text{when } X &= 5, & Y &= 6 + (2 \times 5) = 16 \\ \text{when } X &= 2.5, & Y &= 6 + (2 \times 2.5) = 11 \end{aligned}$$

and similarly for any other value of X .

The general form of this equation is $Y = a + bX$ where a is the value of Y when $X = 0$ and b is the slope of the line. By slope is meant the increase in Y corresponding to one unit of increase in X . Thus it becomes a simple matter to write the equation of any set of points which lie on a straight line.

The problem of measuring trend by a straight line consists of fitting a line to a set of points which are scattered. That is, a straight line cannot be found on which all of the points lie; therefore the task is to determine the position of a line which will describe an average path with respect to the points. The process usually employed in find-

ing such a line is known as the *least squares* method which may be stated as follows: *a line fitted to a time series by the least squares method is that line from which the sum of the squares of the ordinates⁴ of the series will be a minimum.*

Fitting a least-squares line therefore consists of finding values for a and b in the equation, $Y = a + bX$, such that the sum of the squared deviations (ordinates) of the values of the series from the resulting line will be less than the sum of the deviations similarly measured from any other line. The desired values of a and b are obtained by solving simultaneously two normal equations,⁵

$$\Sigma Y = Na + b \Sigma X \quad (1)$$

$$\Sigma (XY) = a \Sigma X + b \Sigma X^2 \quad (2)$$

in which the symbols have the following meanings:

ΣY = the sum of the ordinates measured from zero as an origin.

ΣX = the sum of the abscissas measured from the year preceding the first one of the given series as an origin.

$\Sigma (XY)$ = the sum of the products of corresponding abscissas and ordinates, the ordinates and abscissas being measured as defined above.

ΣX^2 = the sum of the squares of the abscissas as defined above.

a and b = the number of items in the series.

N = the constants to be found in the equation $Y = a + bX$ in which a is the value of Y when $X = 0$, and b is the unit slope of the line, or the change in Y associated with every unit change in X .

The computation of the values to be used in the normal equations is shown in Table 121 for hypothetical data for a five-year period. Substituting the values from the table in equations (1) and (2) gives

$$60 = 5a + 15b \quad (3)$$

$$200 = 15a + 55b \quad (4)$$

$$180 = 15a + 45b \quad \text{multiplying (3) by 3} \quad (5)$$

$$20 = 10b \quad \text{subtracting (5) from (4)} \quad (6)$$

$$b = 2$$

Substituting the value of b in (3)

$$5a = 60 - 30$$

$$a = 6$$

and the equation is $Y = 6 + 2X$ with 1909 as an origin for abscissas. Substituting successive values of X in this equation gives trend values as follows:

⁴ The distances parallel to the Y -axis are used because the line is to be fitted to the ordinates. There is no question of a fit to the abscissas which are spaced equally.

⁵ The derivation of these equations lies outside the scope of this book. For a complete exposition the reader should consult any textbook on probability theory and least squares.

$$\begin{array}{rcl}
 X & bX + a = Y, & \\
 1 & 2 + 6 = 8 & \\
 2 & 4 + 6 = 10 & \\
 3 & 6 + 6 = 12 & \\
 4 & 8 + 6 = 14 & \\
 5 & 10 + 6 = 16 &
 \end{array}$$

TABLE 121
COMPUTATION OF THE VALUES TO BE USED IN THE NORMAL EQUATIONS IN FITTING A
LINE OF LEAST SQUARES

YEAR	1909 ORIGIN				1912 ORIGIN			
	Y	X	XY	X ²	Y	x	xy	x ²
1910.....	8	1	8	1	8	-2	-16	4
1911.....	9	2	18	4	9	-1	-9	1
1912.....	15	3	45	9	15	0	0	0
1913.....	11	4	44	16	11	+1	+11	1
1914.....	17	5	85	25	17	+2	+34	4
Total.....	60	15	200	55	60	0	+20	10

If the origin on the X axis is shifted from 1909 to 1912, the computation of the values for the normal equations changes to the form shown in the second part of Table 121. Substituting these values in the normal equations gives

$$\begin{array}{rcl}
 60 & = & 5a + 0(b) \\
 20 & = & 0(a) + 10b \\
 \text{hence } a & = & 12 \\
 & b & = 2
 \end{array}$$

and the equation becomes $Y = 12 + 2x$ with 1912 as an origin.⁶ The trend values obtained by substituting successive values of x in this equation are identical with those obtained from the equation, $Y = 6 + 2X$ with origin at 1909.

$$\begin{array}{rcl}
 x & bx + a = Y, & \\
 -2 & -4 + 12 = 8 & \\
 -1 & -2 + 12 = 10 & \\
 0 & 0 + 12 = 12 & \\
 +1 & +2 + 12 = 14 & \\
 +2 & +4 + 12 = 16 &
 \end{array}$$

When the central or median year of the series is taken as the origin of abscissas, $\sum x$ will always be zero. Hence the two normal equations become

$$\begin{array}{l}
 \sum Y = Na \\
 \sum (xY) = b\sum x^2
 \end{array}$$

⁶ As explained in chapter XVI, p. 400, footnote 10, capital letters are used to denote the values of a variable measured from the origin or beginning of the scale as zero and small letters are used to denote deviations from the average of the variable as zero.

This is the most convenient form in which to use them since the unknown terms a and b are separated in the two equations.⁷

In subsequent work it will be necessary to have a value of the trend for every value of the original data. This will mean many substitutions in the trend equation in dealing with a long series of data. The work can be shortened by writing the arithmetic average of Y as the trend value for the middle year and adding and subtracting the slope in the two directions from the average. The work can be performed speedily on either a calculating machine or an adding machine.

When the series contains an even number of years as in Table 122 the middle time period falls midway between two years. Each of these years, therefore, is one-half of a time interval from the center. Thus in the table the x value of 1925 is taken as $-.5$ and that of 1926 as $+.5$. The other years have x values as indicated in the table. The computation of the xY column is unchanged. If a number of years is included in the series, it will be found advantageous to obtain the sum of the x^2 by formula. When time is measured at equal intervals on the x axis from the center of the series as an origin, the general formula is

$$\Sigma x^2 = \frac{N^3 - N}{12}$$

in which N is the number of years or time periods included in the data.

For the example of Table 122, $\Sigma x^2 = \frac{10^3 - 10}{12} = 82.5$ and the trend increment equals $353.2 \div 82.5 = \$4.28$ per year.

The computation can be changed to avoid dealing with the fractional values of x by using $2x$ instead of x as shown in Part B of Table 122.⁸ The sum of the third column gives $\Sigma(2xY)$; hence it must be divided by 2 to get the value to be used in the numerator in computing the trend increment. If the value of Σx^2 is computed by squaring the individual $2x$'s instead of by the formula, the sum obtained by adding the column $(2x)^2$ must be divided by 4 to obtain Σx^2 .

The values of the trend are written in the last column of the table. One-half of the annual increment is added to the average (41.61) to obtain the trend value for 1926 and one-half of the annual increment

⁷ The shift from 1909 as an origin to 1912 can be made algebraically by translation of the Y -axis. In the equation $Y = 6 + 2X$ with 1909 as an origin, substitute $X = x + 3$. Then $Y = 6 + 2(x + 3) = 6 + 2x + 6 = 12 + 2x$, the same as was obtained by substituting in the normal equations.

⁸ The change amounts to using one-half year as the time interval instead of the full year interval of Part A.

TABLE 122

COMPUTATION OF STRAIGHT-LINE TREND FOR EARNINGS PER SHARE OF THE CHICAGO AND NORTHWESTERN RAILROAD COMPANY, 1921-30*

YEAR	A			B			
	EARNINGS PER SHARE Y	x	xY	Y	2x	2xY	TREND Y _t
1921.....	\$ - 5.56†	- 4.5	+ 25.020	\$ - 5.56	- 9	+ 50.04	22.35
1922.....	39.73	- 3.5	- 139.055	39.73	- 7	- 278.11	26.63
1923.....	39.01	- 2.5	- 97.525	39.01	- 5	- 195.05	30.91
1924.....	34.25	- 1.5	- 51.375	34.25	- 3	- 102.75	35.19
1925.....	48.15	- .5	- 24.075	48.15	- 1	- 48.15	39.47
1926.....	55.46	+ .5	+ 27.730	55.46	+ 1	+ 55.46	43.75
1927.....	44.32	+ 1.5	+ 66.480	44.32	+ 3	+ 132.96	48.03
1928.....	53.84	+ 2.5	+ 134.600	53.84	+ 5	+ 269.20	52.31
1929.....	69.65	+ 3.5	+ 243.775	69.65	+ 7	+ 487.55	56.59
1930.....	37.25	+ 4.5	+ 167.625	37.25	+ 9	+ 335.25	60.87
	10)416.10		+ 665.230			+ 1,330.46	
	41.61		- 312.030			- 624.06	
			+ 353.200			2)706.40	
						353.2	

$$\text{trend increment} = \frac{353.2}{82.5} = 4.28$$

$$\text{half-trend increment} = \frac{4.28}{2} = 2.14$$

* Irving Bussing, *Railroad Debt Reduction* (New York: Savings Banks Trust Company, 1937), p. 45.
† Deficit.

is subtracted to obtain the value for 1925. Trend values for years earlier than 1925 are obtained by subtracting the annual increment, 4.28 successively. Similarly values for years after 1926 are obtained by adding the increment. If all of these operations have been performed correctly, the average of the two end trend values will be equal to the initial trend value at the center of the time period (arithmetic average of the original data). Thus $(22.35 + 60.87) \div 2 = 41.61$.

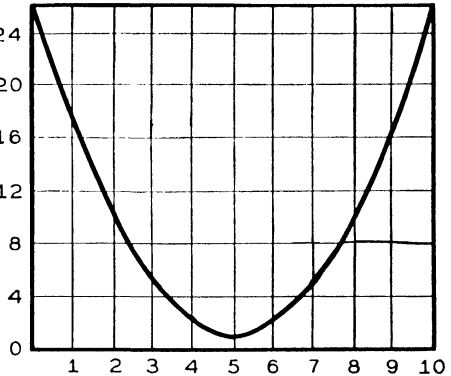
The earnings per share of the Chicago and Northwestern Railroad Company have been selected in order to show the effect of a negative item in computing straight-line trend. In the first row of Table 122 the deficit of \$5.56 multiplied by an x value of -4.5 gives a positive (xY) of 25.020, whereas with an increasing trend a negative value would ordinarily occur in this position. This positive (xY) value increases the slope of the trend line which is exactly the effect that should be obtained from the 1921 deficit in earnings.

A decreasing trend requires no change in the method of computation as given. A negative value for $\Sigma(xY)$ shows that the trend is decreasing since Σx^2 must always be positive. A negative increment,

or properly speaking a decrement, is used in the same way as a positive slope, but taking account of the negative sign results in subtracting the decrement for years following the middle year and adding it for years prior to the middle one.

Other Mathematical Functions.—Many series have trend components which cannot be appropriately represented by straight lines but require the use of more complicated equations. There is no limit to the types of functions which are available, but only the parabola and the logarithmic straight line are used commonly enough to warrant separate discussion in this book.⁹

Parabola: The parabola is one of a whole family of curves of a shape such that they can be intersected at two points by a straight line. The equations which represent them are known as equations of the second degree and are distinguished by the appearance in the equation of one or both of the variables to the second power. The accompanying diagram shows the shape of the parabola and one of the many positions in which it may appear. A particular problem may require the segment of the curve between $X=3$ and $X=5$; another may require the segment between $X=4$ and $X=7$. The conditions of any given series of data will determine automatically what segment of the parabola will be fitted.



The equation of the parabola which is useful in statistical work is the form: $Y = a + bX + cX^2$. This equation contains the three unknown coefficients, a , b , and c , which must be determined for a particular series of data by the least-squares method in order to obtain the best fitting parabola for the series. The three normal equations to be solved are

$$\begin{aligned}\Sigma Y &= Na + b\Sigma X + c\Sigma X^2 \\ \Sigma(XY) &= a\Sigma X + b\Sigma X^2 + c\Sigma X^3 \\ \Sigma(X^2Y) &= a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4\end{aligned}$$

When the median year of the series is used as the origin of the X variable, Σx and Σx^3 become zero, because the positive and negative

⁹ For a very complete explanation of the various families of curves see T. R. Running, *Empirical Formulas* (New York: John Wiley & Son, 1917).

values of x cancel in pairs. The equations become

$$\begin{aligned}\Sigma Y &= Na + c\Sigma x^2 \\ \Sigma (xY) &= b\Sigma x^2 \\ \Sigma (x^2Y) &= a\Sigma x^2 + c\Sigma x^4\end{aligned}$$

The value of b can be obtained immediately and the values of a and c can be obtained by solving the first and third equations simultaneously.

The value of a obtained from solving these equations has the same meaning as in the straight line, i.e., the value of the curve at the origin, or in this case the middle year. It is not the value of the arithmetic average of the data as in the straight line. Also the value of b is not the slope of the parabola since it has a changing slope. The value of c , of course, has no counterpart in the straight line.

TABLE 123
COMPUTATION OF PARABOLIC TREND FITTED TO POSTAL RECEIPTS AT BUFFALO, NEW YORK, 1920-33*

YEAR	POSTAL RECEIPTS (thousand dollars) Y	$2x$	$2xY$	$4x^2Y$	TREND Y_t
1920.....	3,490	- 13	- 45,370	589,810	3,234
1921.....	3,499	- 11	- 38,489	423,379	3,684
1922.....	3,928	- 9	- 35,352	318,168	4,062
1923.....	4,385	- 7	- 30,695	214,865	4,370
1924.....	4,506	- 5	- 22,530	112,650	4,607
1925.....	4,874	- 3	- 14,622	43,866	4,772
1926.....	4,901	- 1	- 4,901	4,901	4,867
1927.....	4,785	+ 1	+ 4,785	4,785	4,891
1928.....	4,747	+ 3	+ 14,241	42,723	4,843
1929.....	4,888	+ 5	+ 24,440	122,200	4,724
1930.....	4,755	+ 7	+ 33,285	232,995	4,535
1931.....	4,342	+ 9	+ 39,078	351,702	4,274
1932.....	3,720	+ 11	+ 40,920	450,120	3,942
1933.....	3,532	+ 13	+ 45,916	596,908	3,540
	60,352		2) +10,706 + 5,353	4) 3,509,072 877,268	

$$\Sigma x^2 = \frac{N^2 - N}{12} = \frac{(14)^2 - 14}{12} = 227.5$$

$$\Sigma x^4 = \Sigma x^2 \left(\frac{3N^2 - 7}{20} \right) = 227.5 \times 29.05 = 6,608.875$$

Normal Equations

$$\begin{aligned}\Sigma Y &= Na + c\Sigma x^2 \\ \Sigma (xY) &= b\Sigma x^2 \\ \Sigma (x^2Y) &= a\Sigma x^2 + c\Sigma x^4\end{aligned}$$

from which

$$\begin{aligned}a &= 4,887.73 \\ b &= 23.53 \\ c &= -35.53\end{aligned}$$

Substituting

and the equation is

$$60,352 = 14a + 227.5c$$

$$5,353 = 227.5b$$

$$877,268 = 227.5a + 6,608.875c$$

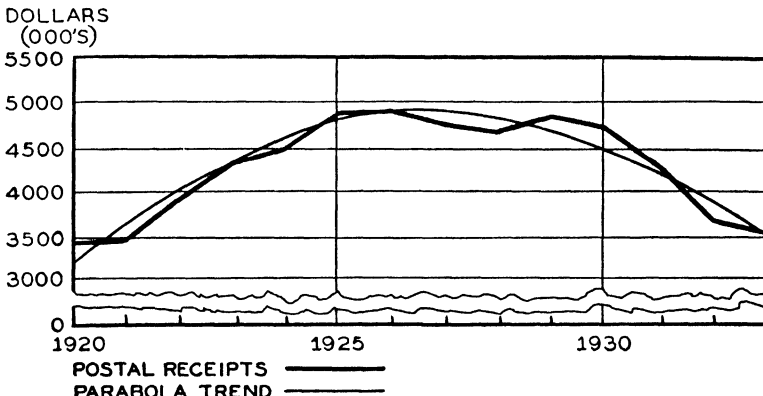
$$Y = 4,887.73 + 23.53x - 35.53x^2$$

* Mimeographed *Statement of Postal Receipts at Fifty Selected Offices*, released monthly by United States Post Office Department.

The computations necessary to obtain a parabola fitted to Buffalo postal receipts are given in Table 123. The equation is $Y = 4,887.73 + 23.53x - 35.53x^2$ (postal receipts in thousands of dollars; x origin at January 1, 1927) and the trend values obtained by substituting successive values of x in the equation are shown in the last column of the table. The original data and the parabolic trend are plotted in Figure 84.

FIGURE 84

PARABOLIC TREND FITTED TO POSTAL RECEIPTS AT BUFFALO, NEW YORK 1920-33



Data from Table 123.

From the mathematical point of view the parabola is an excellent fit for postal receipts. From the statistical point of view it is unsatisfactory. These data were following a definite law of expansion through 1925; they neither expanded nor contracted from 1926 through 1929 but contracted sharply after 1929. This decline was partly cyclical and partly due to the atrophying of mail-order business that had been very large in earlier years. In the light of these facts the parabolic trend is satisfactory as far as 1929 or 1930, but fails to distinguish cyclical movements from trend after 1930. This example leads to the conclusion that knowledge of the economic setting of a series is as important as ability to fit mathematical curves. In this case the curve follows the data too closely. In another the curve may fail to follow closely enough. In general, care must be taken to supplement knowledge of mathematical methods with an adequate comprehension of the circumstances surrounding a series of data.

Logarithmic straight line: The logarithmic line should be used when a preliminary plot of the data on a semi-logarithmic scale shows

the growth component to be approximately a straight line. That means, of course, that the trend is changing at a constant relative rate. The equation is

$$\log Y = a + bx$$

and the normal equations used in determining the values of a and b are,

$$\begin{aligned}\Sigma \log Y &= Na \\ \Sigma (x \log Y) &= b \Sigma x^2\end{aligned}$$

with x origin at the middle of the period.

The production of wood pulp in the United States by the sulphate process provides an example of this type of trend. The detailed computation is given in Table 124. The original data and the trend are plotted in Figure 85. In order to afford a direct comparison between the logarithmic line and a parabola, the latter was also computed and is shown on the chart. The two can be most easily distinguished during the last three years.

FIGURE 85
LOGARITHMIC TREND FITTED TO PRODUCTION OF WOOD PULP, 1923-37

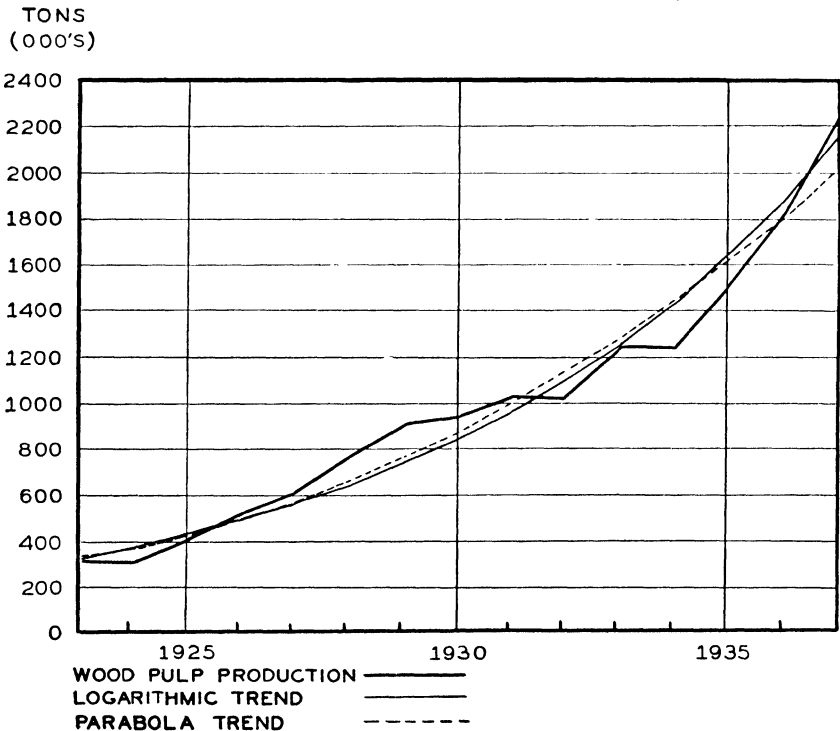


TABLE 124

COMPUTATION OF LOGARITHMIC LINE OF TREND FITTED TO PRODUCTION OF
WOOD PULP BY THE SULPHATE PROCESS IN THE UNITED STATES, 1923-37

YEAR	PRODUCTION OF SULPHATE PULP* (thousand tons) Y	x	$\log Y$	$x \log Y$	TREND IN LOGARITHMS $Y \log t$	TREND IN NATURAL NUMBERS Y_t
1923	312	-7	2.49415	-17.45905	2.52179	332
1924	303	-6	2.48144	-14.88864	2.57979	380
1925	410	-5	2.61278	-13.06390	2.63779	434
1926	520	-4	2.71600	-10.86400	2.69579	496
1927	603	-3	2.78032	-8.34096	2.75379	567
1928	774	-2	2.88874	-5.77748	2.81179	648
1929	911	-1	2.95952	-2.95952	2.86979	741
1930	950	0	2.97772		2.92779	847
1931	1,033	+1	3.01410	+3.01410	2.98579	968
1932	1,029	+2	3.01242	+6.02484	3.04379	1,106
1933	1,259	+3	3.10003	+9.30009	3.10179	1,264
1934	1,246	+4	3.09552	+12.38208	3.15979	1,445
1935	1,508	+5	3.17840	+15.89200	3.21779	1,651
1936	1,817	+6	3.25935	+19.55610	3.27579	1,887
1937	2,220	+7	3.34635	+23.42445	3.33379	2,157
			43.91684	+89.59366 -73.35355 +16.24011		

$$\Sigma x^2 = \frac{15^2 - 15}{12} = 280$$

Normal equations

$$\begin{aligned}\Sigma(\log Y) &= Na \\ \Sigma(x \log Y) &= b \Sigma x^2\end{aligned}$$

Substituting

$$\begin{aligned}15a &= 43.91684 \\ 230b &= 16.24011 \\ a &= 2.92779 \\ b &= .05800\end{aligned}$$

$$\log Y = 2.92779 + .05800x$$

* *Survey of Current Business* (Annual Supplement, 1938), p. 142.

From 1923 to 1935 the two trends are so close that it would make little difference in practice which was used. During the last two years, however, the logarithmic line is appreciably higher than the parabola. This difference is the result of the fact that the logarithmic line is increasing at a constant relative rate whereas the parabola is increasing at less than a constant relative rate and the variation becomes most prominent at the upper end of the range.

The important question is, which of the two provides the better trend for the data. The answer depends entirely upon the observer's judgment and therefore is subject to correction in future years. For the period given, the parabola describes a path which measures cycles slightly more in agreement with the general movement of business. On the other hand, if production of sulphate pulp should continue to

expand at a constant rate in future years, the logarithmic line would be preferable for the given period as well as in future years. The rate of growth during these years has been in excess of 14 per cent annually. It does not seem that this extremely high rate can be continued in future years; hence the conservative point of view would indicate the use of the parabola and the interpretation of the production in 1937 as a high amplitude positive cycle.

Advanced curves: Occasionally other curves are used in fitting a trend. It will be sufficient here to list some of the curves that are developed in more detail in the references at the end of the chapter.

1. The algebraic equation of the third or higher degrees,

$$Y = a + bX + cX^2 + dX^3 + \dots + lX^n.$$

2. The logarithmic parabola,

$$\log Y = a + bX + cX^2.$$

3. The hyperbola,

$$Y = \frac{a}{X} + c.$$

4. The Gompertz curve,

$$Y = ab^{e^x}.$$

The small amount of space devoted to these curves is not to be taken as evidence of their lack of value. They are not treated in detail because examples which require their use occur infrequently.

Moving Trend.—The moving-trend method combines the features of a moving average and a mathematical equation. The moving trend can change direction with the data as is the case with a moving average, but the moving trend can be computed for current data since the moving element is a mathematical equation (usually a straight line) instead of an average. The method is therefore especially useful in continuous computations involved in the analysis of current data as they appear.

A moving trend is also advantageous for a series of data which has experienced a marked change of direction during any period of years that is being studied. It is used in the example in chapter XXIV.

WHY TREND IS MEASURED

Thus far in the chapter attention has been devoted exclusively to the discussion of the location of trend and its measurement. A question

must now be raised as to why trend should be measured. The two answers are (1) to remove the trend component from a series and (2) to study the trend component itself free from the disturbing influences of other components.

The Removal of Trend

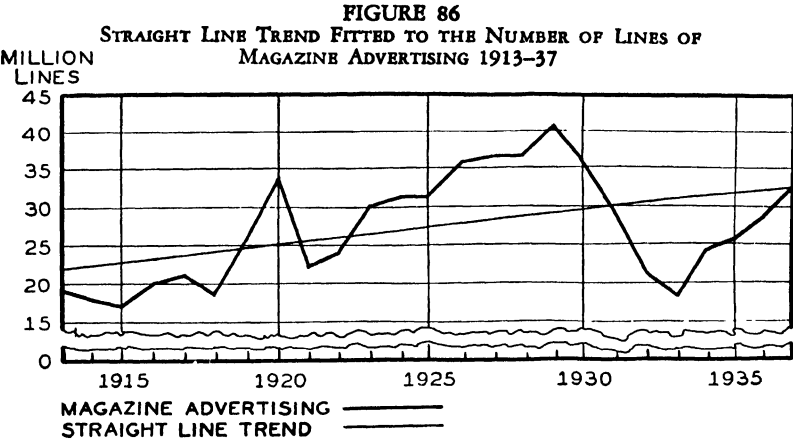
In a series of annual data,¹⁰ the preliminary analysis consisting of the adjustment for calendar variation and the correction for price change having been carried out, there remain in the series trend and cyclical fluctuation, the latter including any unpredictable irregularities that may be present. The trend could be removed by subtracting each trend value from the corresponding item of original data. This step, $Y - Y_t$, would give a set of remainders or residuals with positive or negative signs and a net sum of zero. However, with an increasing trend these residuals would tend to be larger near the end of a given period and smaller near the beginning. With a decreasing trend, the reverse tendency would be found. But this tendency of the residuals to increase or decrease with the trend can be eliminated by dividing each residual by the corresponding trend value. The computation, $\frac{Y - Y_t}{Y_t}$, gives a set of relative residuals which are free from the influence of trend and which can be expressed as per cents above and below zero. But

$$\frac{Y - Y_t}{Y_t} = \frac{Y}{Y_t} - 1 = \frac{Y}{Y_t} - 100\% \text{ and } \frac{Y - Y_t}{Y_t} + 100\% = \frac{Y}{Y_t}.$$

Hence all of this can be accomplished by the single step of dividing each item of original data by the corresponding trend value to obtain a set of relative residuals expressed as per cents above and below 100 per cent.

The removal of trend from lineage of magazine advertising in the period 1913 to 1937 is carried out in Table 125. The series is plotted in Figure 86. A straight line has been fitted in the table for illustrative purposes, but it is not a good fit at the beginning of the period. The trend values are written in column 4 and the relative cycles in column 5. These cycles are plotted in Figure 87. As contrasted with the volume of advertising since World War I, the lineage for the years

¹⁰ The methods of dealing with data recorded at more frequent intervals than once a year are developed in the next chapter.



Data from Table 125.

TABLE 125
COMPUTATION OF RELATIVE CYCLES FROM STRAIGHT-LINE TREND OF
NUMBER OF LINES OF MAGAZINE ADVERTISING IN THE UNITED STATES, 1913-37

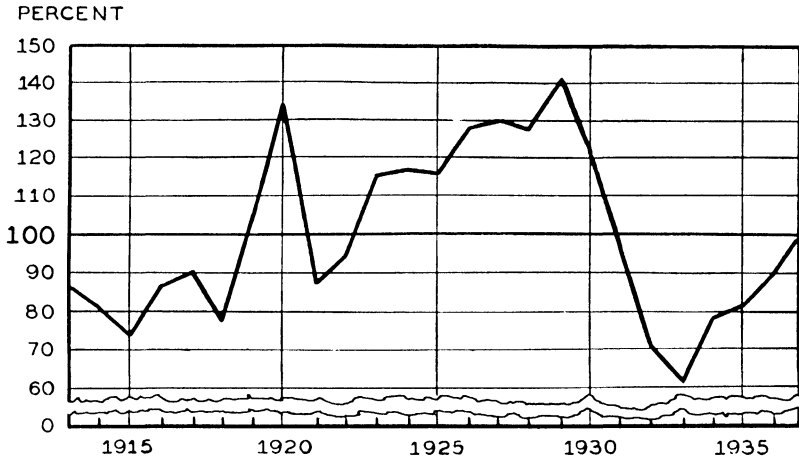
YEAR	(1) MAGAZINE ADVERTISING* (million lines) <i>Y</i>	(2) <i>x</i>	(3) <i>xY</i>	(4) TREND <i>Y_t</i>	(5) RELATIVE CYCLES $\frac{Y}{Y_t}$
1913	19.26	— 12	— 231.12	22.05	87.3
1914	18.29	— 11	— 201.19	22.47	81.4
1915	16.88	— 10	— 168.80	22.89	73.7
1916	20.03	— 9	— 180.27	23.31	85.9
1917	21.26	— 8	— 170.08	23.73	89.6
1918	18.56	— 7	— 129.92	24.15	76.9
1919	25.70	— 6	— 154.20	24.57	104.6
1920	33.64	— 5	— 168.20	24.99	134.6
1921	22.27	— 4	— 89.08	25.41	87.6
1922	24.36	— 3	— 73.08	25.83	94.3
1923	30.24	— 2	— 60.48	26.25	115.2
1924	31.44	— 1	— 31.44	26.67	117.9
1925	31.48	0	0	27.09	116.2
1926	35.50	+ 1	+ 35.50	27.51	129.0
1927	36.46	+ 2	+ 72.92	27.93	130.5
1928	36.38	+ 3	+ 109.14	28.35	128.3
1929	40.61	+ 4	+ 162.44	28.77	141.2
1930	35.81	+ 5	+ 179.05	29.19	122.7
1931	28.91	+ 6	+ 173.46	29.61	97.6
1932	21.16	+ 7	+ 148.12	30.03	70.5
1933	18.66	+ 8	+ 149.28	30.45	61.3
1934	24.32	+ 9	+ 218.88	30.87	78.8
1935	25.38	+ 10	+ 253.80	31.29	81.1
1936	28.54	+ 11	+ 313.94	31.71	90.0
1937	32.05	+ 12	+ 384.60	32.13	99.8
	25) 677.19		+ 2201.13		
	27.09		— 1657.86		
			+ 543.27		

$\Sigma x^2 = \frac{25^2 - 25}{12} = 1300$ trend increment = $\frac{+543.27}{1300} = +.42$

* *Survey of Current Business* (Annual Supplement, 1938), p. 25.

prior to 1919 was small; hence it is shown in Figure 87 as a negative phase of the cycle. This does not appear to be an adequate description of the early period, but there is no alternative so long as a straight-line trend is used.

FIGURE 87
RELATIVE CYCLES OF MAGAZINE ADVERTISING 1913-37



Data from Table 125.

The cyclical movements of magazine advertising as portrayed in Figure 87 could be described as follows. The curve declined from 1913 to 1915, rose during the following two years, declined in 1918, and crossed the average line¹¹ in 1919. Advertising expanded sharply in 1920 but dropped below the average in 1921 and 1922. Consistent expansion followed, reaching a peak in 1929. In response to the depression, magazine advertising dropped sharply after 1929 to a low point in 1933 which was 61 per cent of the average volume. The expansion since 1933 has brought the curve up to the average line in 1937.

¹¹ In describing cycles the trend line serves as a standard or average condition of the data. It is an axis above and below which the alternations of prosperity and depression carry the actual level of activity. It will be convenient to refer to the trend as the average level of the data. This usage requires the reader to think in terms of an average condition which is changing with the passing of time. There is no difficulty involved in this concept once it is clear that what is called above the average in a series of data at one period of time may be less in amount than that which is designated as below the average at a later period of time when dealing with an increasing trend. For example, 29 million lines of magazine advertising in 1936 is considered to be 10 per cent below average, while 26 million lines in 1919 is considered to be 5 per cent above average.

The trend is often referred to as the normal condition of the data. The word "normal" implies the existence of more stability in business conditions than is warranted. It seems preferable, therefore, to abandon "normal" in favor of "average." See chapter XIX, pp. 483-84. for further discussion of this problem.

This statement of the cycles of magazine advertising is, of course, entirely dependent upon the selection of a straight-line trend. Had some other trend been chosen the results would have been somewhat different. This is the situation which always confronts the statistician in fitting a trend. There are no objective tests which can be applied to determine whether the proper trend has been fitted. The preliminary judgment must be made when the trend is selected. The final judgment comes from the study of the resulting cycles. The question is, Are the cycles reasonable? As previously stated a straight line does not give a good description of the pre-war cycles and some questions could be raised concerning later years. The most important test comes for 1937. The cycle stands at 100 per cent, an indication that magazine advertising was neither prosperous nor depressed. This appears to be a reasonable description of the 1937 situation.

Consideration of the Trend Component Itself

The trend may be fitted to a set of data in order to obtain the historical record of growth in the past and up to the present time, or it may be fitted to a past period to obtain a basis for predicting the trend of the data in future years. These two are usually referred to as historical trend and trend for projection. When dealing with historical trend the actual data to which the fit is to be made are used as the basis for judging what type of curve to use. The problem is solely that of obtaining the best description of the growth component during the period under consideration. When the additional problem of projection is brought in, the trend must be selected so that it will give a good historical description as well as a projection into future years which has a reasonable chance of according with events as they occur. Not infrequently a curve which will serve the historical purpose becomes inadequate when projected. It follows that the most difficult problem of trend fitting is that of finding a curve which will serve the dual purpose of historical record and projection.

What Types of Trend Can Be Projected

Three methods of measuring trend were discussed—moving average, mathematical functions, and moving trend.

The moving average cannot be used for projection. It can be kept abreast of current data with the aid of supplementary free-hand methods only; hence projection is not feasible. If the purpose of computing a trend is to be able to project it, a moving average should not be used.

Sometimes mathematical curves can be projected with satisfactory results, but frequently their projections become absurd. In every case the trend has been fitted to a historical period according to some equation. Projection of the curve assumes that the same equation will hold in the future. How well the facts will accord with the assumption depends upon the extent of the instability of business conditions during the period of projection. If business is in a relatively quiescent state, the direction and intensity of the trend component are likely to be undisturbed. Under these conditions a well-selected trend equation can be projected with satisfactory results. If, on the other hand, business conditions are in a disturbed state such as that witnessed since 1930, the projection of mathematical curves is likely to lead to results which are of little value. Since projection of trend is most valuable to business management in periods of stress, the inflexibility of mathematical curves is a serious drawback to their use for the purpose.

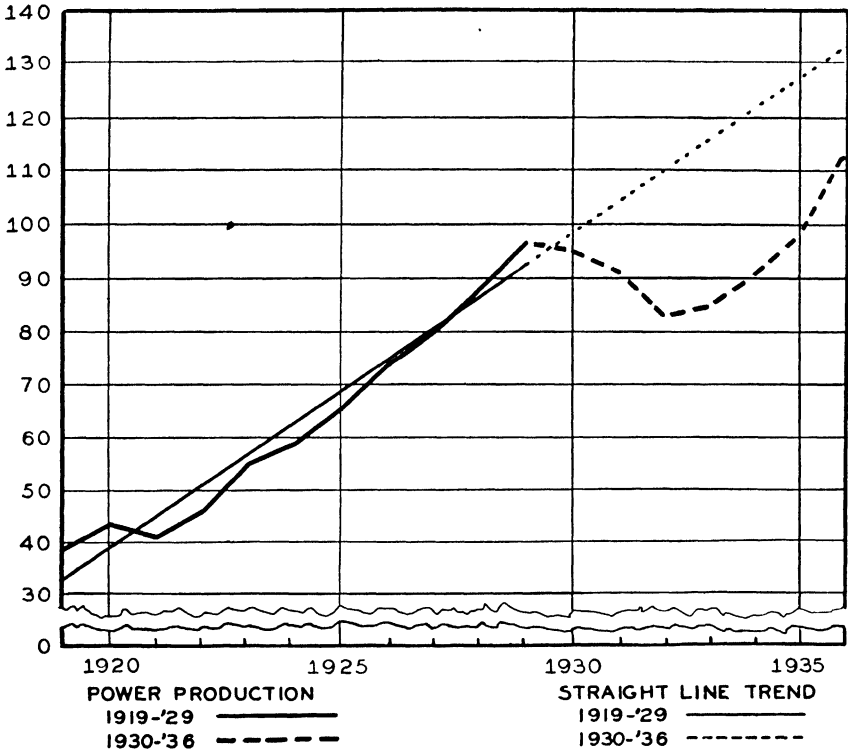
An example in which a straight line is projected will show what may happen. Figure 88 shows the production of electric power from 1919 to 1929. A straight line has been fitted to the eleven-year period and as a historical trend for that period is fully justified. Suppose, however, that the trend had been projected from 1929 to 1936 as a basis for determining the average amount of producing equipment needed to meet the demands for current during the seven-year period. The projected trend is shown on the chart as a dotted line. The actual production is also shown as a dashed line. The actual production stood at about 75 per cent of the projected trend from 1932 to 1935 and rose to 85 per cent in 1936. If equipment had been provided according to the projected trend, only 85 per cent of the capacity would have been needed in 1936 and it is uncertain when, if ever, actual production would overtake this projected trend. To be sure, such a program of equipment expansion, even if undertaken after 1929, would have been abandoned long before 1936. The example is intended merely to show that projection of a straight-line trend in 1929 would not have been desirable.

The difficulty encountered in this example is not peculiar to the use of straight-line trend. It is related to the inflexibility of mathematical functions. It might be argued that some other type of curve would have been better than a straight line in this case, but there is no assurance that it would give as good results in the next case. As a further example consider the parabola fitted to postal receipts in Figure 84.

FIGURE 88

STRAIGHT LINE TREND FITTED TO ELECTRIC POWER PRODUCTION 1919-29
(DATA AND PROJECTED TREND SHOWN 1930-36)

BILLION
K.W.H.



Data from Statistical Abstract, 1937.

For the period 1920 to 1933 the parabola is an excellent fit, but it could not be projected because it would predict the disappearance of postal receipts in a few years. The projection of such a trend is, of course, absurd, yet the difference between this case and others involving projection of mathematical curves is frequently only the degree of absurdity. At the very least we can say that great care should be exercised in the projection of mathematical curves.

A moving trend¹² is specially designed for projection into current periods as the data become available. Projection in advance of current data for limited periods can also be made by assuming data for such periods and computing temporary trend values which, of course, must

¹² The computation of the moving trend is explained in chapter XXIV.

be corrected as the actual data become available. This process is less subject to error than might at first glance be supposed, since the projected values of the trend depend largely on actual values of the data for earlier years included in the base period.

If a moving trend with an eleven-year base had been used for electric power production, the projection for the years 1930, 1931, and 1932, for example, might have been carried out on the assumption that power production during the three years remained at the 1929 amount. The resulting projected trend would, of course, have been somewhat too high, but certainly would have been superior to the projected straight line of the chart.

SUMMARY

The previous discussion has shown that the greatest problem connected with trend fitting is the determination of the proper trend for a given series. The problem attains its maximum complexity when the trend must serve the dual purpose of historical description and projection. Experience and judgment form the necessary background for proper selection. The first step is to construct a graph of the data to which the trend is to be fitted and study the shape of the curve and outline free hand the approximate position of a trend which will give a reasonable description of the cycles present in the particular series. Then determine what type of mathematical equation, if any, will correspond to the indicated trend. In some cases a straight line will be satisfactory; in others a more complicated curve will be required. If the free-hand trend changes direction one or more times during the period being studied, a moving average will probably give the preferable trend.

When none of the preceding alternatives is satisfactory and particularly if the trend is to be computed for current data as they become available, the moving-trend method may prove preferable to any of the others. Its greatest advantage lies in the ease of obtaining current trend values without any change in the values for earlier periods.

PROBLEMS

1. What are the two major problems encountered in dealing with trend?
2. What is meant by a flexible trend?

- 3. Fit a five-year moving average and a fourteen-year moving average to the data of Table 119, column 1, page 562. Compare your results with those given in column 3 of the table. What principles of the use of moving averages are illustrated by this problem?
- 4. Collect the data necessary to bring Table 120, page 567, up to date. Extend the computation of the eight-year moving average through these subsequent years. Criticize the free-hand trend figures of column 5.
- 5. The following are series of hypothetical data to be used for practice in fitting mathematical curves.

YEAR	A	B	C
1st	3	17	25
2d	3	13	33
3d	7	11	29
4th	5	10	37
5th	7	10	46
6th	..	5	..

- a) Compute the equation of a straight line fitted to series A: (1) using the year preceding the 1st year as X origin, (2) using the middle of the time period as x origin.
 - b) Perform the same operations for series B.
 - c) Perform the same operations for series C, but fitting a parabola instead of a straight line.
 - d) Verify the identity of the trend values obtained from the two equations in each of (a), (b), and (c).
6. The consumption of newsprint paper in the United States and the imports of newsprint paper were as follows (in thousand tons):

YEAR	CONSUMPTION*	IMPORTS	YEAR	CONSUMPTION*	IMPORTS
1928	2746	2157	1934	2477	2210
1929	2937	2423	1935	2663	2383
1930	2819	2280	1936	2939	2752
1931	2618	2067	1937	2956	3317
1932	2255	1792	1938	2653	2275
1933	2146	1794	1939	2735	2616

*The consumption data represent approximately 80 per cent of the total newsprint used in the United States.

- a) Fit straight-line trends to the two series.
 - b) Is there evidence here of increasing dependence upon imported newsprint paper?
 - c) Collect data for years since 1939. Do the more recent data indicate that the computed trends should or should not be projected?
7. Fit a logarithmic parabola to the consumption data of Problem 6. Which of the two trends is preferable for this series?

8. What is the twofold purpose of measuring trend?
9. What are the merits and defects of expressing cycles, respectively, in the original units and as relatives of the trend?
10. What type of trend, if any, would be superior to a straight line to measure the trend of magazine advertising in Table 125, page 584? Discuss.
11. What difficulties may be encountered in projecting trends derived by mathematical equations?

REFERENCES

CROXTON, FREDERICK E., and COWDEN, DUDLEY J., *Applied General Statistics*. New York: Prentice-Hall, Inc., 1939.

Chapter XVI contains a detailed explanation of the methods of fitting a variety of mathematical equations to data.

KING, WILLFORD I., "Principles Underlying The Isolation of Cycles and Trends," *Journal of the American Statistical Association*, Vol. XIX, (December, 1924).

KUZNETS, SIMON, "On The Analysis of Time Series," *Journal of the American Statistical Association*, Vol. XXIII (December, 1928).

KUZNETS, SIMON, *Secular Movements in Production and Prices*. Boston: Houghton Mifflin Co., 1930.

A painstaking, logical analysis of the trend problem.

MACAULAY, FREDERICK R., *The Smoothing of Time Series*. New York: National Bureau of Economic Research, 1931.

Chapters I and II are particularly pertinent as collateral to the text, but the entire book should be read.

MITCHELL, WESLEY C., *Business Cycles, The Problem and Its Setting*. New York: National Bureau of Economic Research, 1927.

A thought-provoking discussion of types of trend and the interpretation of a fitted trend, pp. 212-33.

SAZULY, MAX, *Trend Analysis of Statistics*. Washington: Brookings Institution, 1934.

An advanced mathematical treatment of the subject.

CHAPTER XXIII

SEASONAL AND CYCLICAL MOVEMENTS

INTRODUCTION

THIS chapter is devoted to the further development of the methods of analyzing time series. Having discussed the preliminary steps and the measurement of the trend component, there remains the task of studying the rhythmic movements or more specifically the separation of the seasonal variation from the cyclical fluctuation. It was stated in chapter XXI, page 550, that the cyclical component is the residual element of the analysis, i.e., what is left after the other parts of the series have been measured and removed. The principal remaining task therefore is to explain the nature of seasonal variation and the methods of measuring and removing it.

In the preceding chapter the development was in terms of annual data. In seasonal analysis several observations are needed during each year and in practice monthly data are usually employed. Likewise the determination of the seasonal rhythm requires a fairly long period of years. Hence the examples in this chapter will necessarily involve some ten years or more of monthly data. This means that a great amount of computation is inevitable in dealing with seasonal variation.

THE NATURE OF SEASONAL VARIATION

The causes of the seasonal rhythm and its repetitive character have been explained in an earlier chapter. At this point interest will be focused on the further description of seasonal movements according to (1) amplitude, (2) regularity.

Amplitude

By amplitude we mean the extent to which seasonal influences draw the data either above or below the path which they would follow if only the trend and cyclical components were present. The sales of the Woolworth Company were presented in Figure 75, page 546. This is an example of high seasonal amplitude, the peak in December being regularly as much as 60 per cent to 75 per cent above the average level

of the year. Seasonal variations of high amplitude are to be expected in most retail sales figures due to their dependence upon both weather conditions and buying habits related to holidays. Seasonals of high amplitude are also found outside the realm of retail trade in series such as automobile production, postal receipts, cement production, and construction contracts.

There are many other series which exhibit definite seasonal rhythms but of low amplitude. Bank debits for example have a regular repetitive movement, but in August, the lowest month, they are only 11 per cent below the average for the year and in December, the highest month, are only 8 per cent above the average. Other series with low amplitude seasonal are pig-iron production, interest rates on commercial paper, and gross earnings of railroads.

Regularity

All of the series referred to in the preceding section have seasonal movements which recur year after year in the same months and with approximately the same intensity. The indexes of some composite series such as commodity prices and stock prices have no regular seasonal movements. The large number of individual series used in making such indexes may have marked seasonal variations but these cancel out when the individual series are combined in a composite index.

As distinguished from series of this type there are others which exhibit seasonal influences of such irregular character that they are difficult to recognize and even more difficult to measure. A study of the value of building permits in Buffalo, New York, brought out the fact that the entire staff of the permit office might be engaged for several weeks in checking the blue prints of one large project. During that time only purely routine permits were granted and in the interim value declined accordingly. In the following month, perhaps, a large value of permits would be recorded. Briefly put, the operating methods of the permit office led to totally irregular reports on a series that possesses perfectly definite seasonal variations. Other series with irregularity, whatever the cause, are life-insurance sales and call-money rates.

When the statistician encounters irregularity of this sort he makes no attempt to deal with it as seasonal, but rather classifies it as indeterminate and as such allows it to reside with the cyclical fluctuation

THE CONCEPT OF A SEASONAL PATTERN

Definition

*A seasonal variation is a rhythmic movement which recurs annually with approximately the same relative intensity.*¹ This definition places limitations upon the concept of a seasonal variation in three directions and at the same time points the way to methods by which such variation can be measured. First, it recurs annually; all other rhythmic components of a series are eliminated except that regular one with a period of twelve months. Secondly, seasonal variation recurs from year to year with similar intensity; irregular responses to weather factors or man-made conventions are excluded, and only amplitudes that are repeated year after year are included. Finally, relative intensity means that a certain month, say March, is expected to vary in the same direction by the same per cent year after year because of seasonal influence.

All of these requirements may be summarized in the concept of seasonal variation as a pattern which is assumed to be typical of any year of a series. Actually the pattern is obtained in the form of an index composed of 12 monthly relatives whose average is 100. The whole problem of measuring seasonal variation then consists of developing methods of establishing the pattern and its index for a given series.²

Determining the Existence of a Pattern

As pointed out on an earlier page of the chapter a series may contain seasonal swings but unless these seasonal amplitudes recur regularly from year to year the series does not contain a seasonal pattern and no index of seasonal variation can be computed. The first requirement then is to determine that a seasonal pattern exists. In many cases careful study of an ordinary graph of the data will provide the answer. For example, Figure 75, page 546, shows rather clearly that the sales of the F. W. Woolworth Company possess a seasonal pattern. In other series the regularity of seasonal amplitude may be in doubt

¹ This definition is purposely restricted so as to unify the major development of the chapter. In earlier years some methods were proposed that were not expressed in relative form, but these have been superseded in practice. Some series require the use of a changing seasonal, the form of which is discussed at the end of this chapter. In neither of these methods does the seasonal pattern retain the same relative intensity from year to year, yet they are disregarded in the definition and in the major exposition of seasonal variation that follows in order that the basic principles may be presented first.

² This statement has certain exceptions which will be discussed later in the chapter under the head of special problems.

and the pattern, if any, will have to be determined by the use of an auxiliary graph.

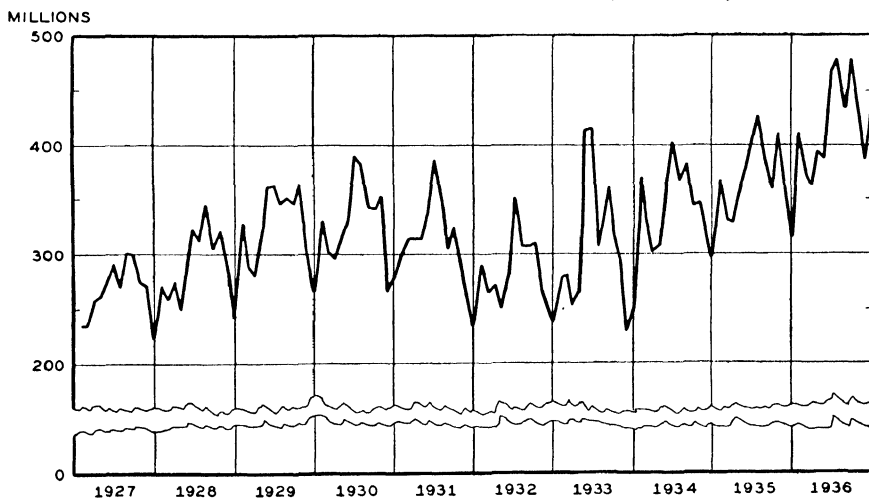
This graph can be explained in a preliminary way as a plot in which the horizontal scale consists of the twelve months of the year and the data for all years are plotted on this common base. However, it is not the original data that are plotted on the vertical scale of this graph but a set of relatives obtained from the original data and containing mainly seasonal influence. Such relatives appear at an early stage of each of the methods of measuring seasonal variation but they are in slightly different form in the several methods. Consequently the test for the existence of a seasonal pattern can be explained in detail in connection with the individual methods to greater advantage than at this point.

METHODS OF MEASURING SEASONAL VARIATION

Many methods of measuring seasonal variation are explained in statistical literature.⁸ Three of these, the moving-average method, the link-relative method, and the ratio-to-trend method, plus an approximate method, will be explained in this chapter.

FIGURE 89

DAILY AVERAGE CONSUMPTION OF SMALL CIGARETTES, MONTHLY, 1927-36



Data from Table 126.

⁸ The student desiring to pursue this subject further will find ample material in the *Journal of the American Statistical Association* since 1922.

TABLE 126
MONTHLY CONSUMPTION OF SMALL CIGARETTES (TAX-PAID WITHDRAWALS)
IN THE UNITED STATES, 1927-1936*
(daily averages, 000,000 omitted)

MONTH	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936
Jan.	234.5	270.0	327.7	329.3	302.2	289.1	278.1	370.4	365.7	410.5
Feb.	236.0	259.7	287.9	302.3	315.6	264.8	280.5	327.4	332.4	371.2
Mar.	258.9	273.2	280.3	295.6	316.2	272.5	257.2	301.1	329.0	361.1
Apr.	262.7	250.4	320.3	317.8	315.7	252.1	265.8	309.8	356.6	395.6
May	275.5	286.8	360.3	332.3	337.0	280.2	413.6	360.5	377.7	387.9
June	291.2	323.0	361.3	391.7	383.6	352.0	415.4	401.5	404.0	467.0
July	267.0	313.7	345.9	382.5	345.2	307.5	307.3	366.3	423.8	477.5
Aug.	300.9	342.8	352.6	341.2	307.1	308.4	360.9	381.0	386.3	433.2
Sept.	299.8	304.2	345.0	339.7	323.2	310.4	317.6	343.1	359.1	478.1
Oct.	275.9	320.1	361.4	353.1	288.9	269.4	296.0	345.7	410.0	425.9
Nov.	269.8	284.5	301.4	265.1	261.7	253.8	227.8	324.2	360.0	385.9
Dec.	221.6	242.4	266.5	279.8	235.3	236.1	251.6	297.1	317.5	427.3
Av.	266.2	289.4	326.1	327.7	310.8	283.0	306.2	344.2	368.8	418.6

* *Survey of Current Business* (Annual Supplements, 1932, 1936, and 1938), pp. 170, 101, and 116 respectively (data changed to daily averages).

One series of monthly data, consumption of cigarettes for the ten-year period 1927 to 1936,⁴ will be used to illustrate all four methods of measuring seasonal variations. The daily average consumption of cigarettes in millions is presented in Table 126 and Figure 89. The chart shows the existence of seasonal movement but the exact pattern is not apparent without further study. The details of the analysis of this series by each of the four methods follow.

An Approximate Method

The steps described in this section are not designed to give a precise measure of seasonal variation. The chief virtue of the method is the fact that an approximate seasonal pattern is secured by a very simple process.

The first step is to divide the data for each month of the first year by the average amount for the first year, to divide the data for each month of the second year by the average amount for the second year, and so on to the end of the period. The relatives thus obtained vary from 100 per cent from month to month largely because they contain seasonal movements. If a strong upward trend is present in the data the relatives near the end of the year will be raised somewhat and those near the beginning of the year will be lowered to the same extent.

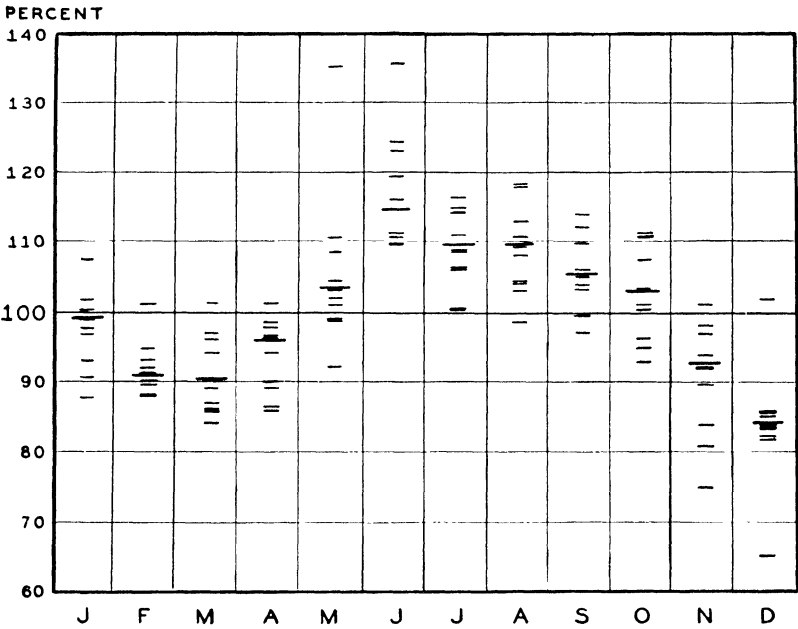
⁴ The data are actually tax-paid withdrawals from manufacturers' warehouses and therefore contain a seasonal movement expressing variations in dealers' stocks rather than variations in consuming habits.

The relatives will exhibit a reverse tendency, if a strong downward trend is present. The presence of pronounced cyclical fluctuation will affect the relatives to a greater extent than trend. It must therefore be understood that the first step of the approximate method accomplishes best the desired purpose of segregating the seasonal movements when the trend and cyclical components of a series produce in any year a change which is small compared with the seasonal movement. At the other extreme the approximate method should not be used when the trend and cyclical components produce more change in the course of a year than the seasonal movements.

The relatives of the first step are given in Table 127. They are in the form suggested in the preceding section, hence the computation of seasonal is interrupted at this point to test for the existence of a seasonal pattern. The relatives of Table 127 are plotted in Figure 90. The ten January relatives are indicated by horizontal dashes in the first vertical section of the chart, the February relatives are represented in the second vertical section, etc., to December. If the series contained identical seasonal movements year after year, and no trend or cycle, all of the dashes in any vertical section would coincide. If the series contained identical seasonal movements and trend but no cycle, the dashes would very nearly coincide because most of the effect of trend is removed by using the average of each year as a base for the relatives of that year. When cyclical fluctuation is present, however, it tends to spread the relatives in any month. For a particular month relatives that measure the depth of depression will be low and relatives that measure the peak of prosperity will be high, but the majority of the relatives will group themselves within a narrow range on the vertical scale, on the assumption that in any given month there will not be more than one or two relatives affected by extreme cyclical movements. The more closely grouped the dashes, the better the pattern is defined. Unless the dashes show definite grouping in a majority of the months, no repetitive seasonal movement is present in the data and a seasonal pattern should not be computed.

Study of the chart shows that some of the months exhibit better grouping than others. Specifically, six of the ten relatives fall close together in January. The February relatives show the best grouping of any month. The March and April relatives are satisfactorily grouped. There is less concentration in May, one relative being markedly higher than the others. In June the five smaller relatives are well grouped

FIGURE 90
APPROXIMATE METHOD: TEST FOR SEASONAL PATTERN OF RELATIVES OF ANNUAL AVERAGES



Data from Table 127.

TABLE 127
APPROXIMATE METHOD: ORIGINAL DATA EXPRESSED AS RELATIVES OF THE ANNUAL AVERAGES. SEASONAL INDEX AT THE RIGHT
CIGARETTE CONSUMPTION DATA FROM TABLE 126

MONTH	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	MEDIAN ^s	SEASONAL INDEX (adjusted medians)
Jan. ...	88.1	93.3	100.5	100.5	97.2	102.2	90.8	107.6	99.2	98.1	98.7	99.3
Feb. ...	88.7	89.7	88.3	92.2	101.5	93.6	91.6	95.1	90.1	88.7	90.9	91.5
Mar. ...	97.3	94.4	86.0	90.2	101.7	96.3	84.0	87.5	89.2	86.3	89.7	90.3
Apr. ...	98.7	86.5	98.2	97.0	101.6	89.1	86.8	90.0	96.7	94.5	95.6	96.2
May ...	103.5	99.1	110.5	101.4	108.4	99.0	135.1	104.7	102.4	92.7	103.0	103.7
June ...	109.4	111.6	110.8	119.5	123.4	124.4	135.7	116.6	109.5	111.6	114.1	114.8
July ...	100.3	108.4	106.1	116.7	111.1	108.7	100.4	106.4	114.9	114.1	108.6	109.3
Aug. ...	113.0	118.5	108.1	104.1	98.8	109.0	117.9	110.7	104.7	103.5	108.6	109.3
Sept. ...	112.6	105.1	105.8	103.7	104.0	109.7	103.7	99.7	97.4	114.2	104.6	105.3
Oct. ...	103.6	110.6	110.8	107.7	93.0	95.2	96.7	100.4	111.2	101.7	102.7	103.4
Nov. ...	101.4	98.3	92.4	80.9	84.2	89.7	74.4	94.2	97.6	92.2	92.3	92.9
Dec. ...	83.2	83.8	81.7	85.4	75.7	83.4	82.2	86.3	86.1	102.1	83.6	84.1

but the five larger ones are scattered. The July, August, and September relatives are moderately well grouped. The October and November relatives are too scattered and December has eight closely grouped rela-

tives with one wide variation in each direction. If several of the months showed as little concentration of the relatives as October and November it would be better to conclude that no seasonal pattern is present. As the chart stands there is enough concentration to warrant the use of a seasonal pattern.

The center or median position of the ten dashes is marked on each vertical section by a dash longer than the others. The shape of the seasonal pattern can be seen by following the heavy dashes across the chart.

The further steps in obtaining an index of seasonal variation by the approximate method consist in finding the median of the relatives of Table 127 for each month and adjusting the values of the 12 medians so that their sum will be 1200. The second step is necessary in order to obtain a seasonal index in which the per cents for the several months are balanced above and below 100 per cent. By using the index in this form the cyclical fluctuations and trend will not be disturbed by the removal of the seasonal variation from a series. This point will be elaborated later in the chapter.

The medians are shown in next to the last column of the table. The median is used in preference to the arithmetic average because it has a tendency to neutralize the effect of extreme items either above or below the point of concentration of the relatives. These monthly medians are then adjusted so that their sum will be 1200. Their sum as computed is 1192.4. The adjusting consists of dividing each median by the average of the medians or by multiplying each median by 1.0064, the reciprocal of the average of the medians. The adjusted medians are given in the last column of the table. This is the seasonal index by the approximate method.

This method cannot be recommended for use in a series containing a steep trend or cycles of high amplitude. It will give fairly good results in a series containing mild trend and cyclical components. The comparison of the results of this method with the three which are described on succeeding pages can be seen in Figure 94.

Moving-Average Method

The use of the moving average to separate seasonal movements from trend and cycle dates back many years, but the form of it now in use was published originally in the *Federal Reserve Bulletin*.⁵ A

⁵ "Index of Production in Selected Basic Industries," *Federal Reserve Bulletin* (December, 1922). Method of measuring seasonal variation explained on pp. 1415-17.

twelve-month moving average is first fitted to the data.⁶ The twelve-month moving average does not contain any seasonal variation since one complete seasonal wave is included in each computation. This twelve-month moving average describes the path which the data would follow if no seasonal movements were present. That is, the seasonal expansion and contraction cancel each other in a twelve-month average. This would be true regardless of the order in which the calendar months appeared in the moving average.

The next step in obtaining a pattern is to divide each monthly item of original data by the corresponding value of the twelve-month moving average. Division is the proper operation at this point rather than subtraction because seasonal variation is defined as a component which repeats itself from year to year with the same relative intensity. The twelve-month moving average may be thought of as describing the path of the trend and cyclical fluctuations combined. Therefore the ratio, original data divided by twelve-month moving average, gives the original data as a percentage of trend and cycle combined. When these ratios are above 100 per cent, positive seasonal movement is the cause; when they are below 100 per cent, negative seasonal movement is the cause. The ratio maintains the proper relation between the seasonal and the trend-cycle base. An illustration may help to clarify this point. Suppose that the sales of a department store had experienced a great expansion during a given ten-year period. The components of December sales might be as follows:

	(1) SALES	(2) TREND	(3) CYCLE	(4) SEASONAL	(5) $1 \div (2+3)$ (per cent)
Dec. of 1st yr.	\$10,000	\$5,000	\$1,000	\$4,000	166 $\frac{2}{3}$
:	:	:	:	:	:
Dec. of 10th yr.	\$1,000,000	\$500,000	\$100,000	\$400,000	166 $\frac{2}{3}$

The computation shows that regardless of dollar value of sales the December sales were 66 $\frac{2}{3}$ per cent above the base (trend plus cycle) because of seasonal business. This is precisely what is meant by a seasonal variation which repeats itself year after year with equal relative intensity.

The computation of a twelve-month moving average requires the

⁶ The twelve-month moving average used here serves a purpose entirely distinct from the use of the moving average to measure trend as explained in the preceding chapter.

same steps that were explained in the preceding chapter⁷ in fitting a moving-average trend with an even number of years in the base. An average of data for the months of January to December inclusive falls at the middle of the year or at July 1, and an average of data for the months of February to the following January, inclusive, falls on August 1. Neither of these averages coincides in time with the July original data which are taken as of July 15.⁸ The value of the moving average coinciding in time with the July original data is obtained by taking the average of the two moving-average figures at July 1 and August 1, respectively. The results of performing these two operations with the data for cigarette consumption of Table 126 are given in Table 128. The intermediate steps in the computation of the moving average have been omitted because of lack of space.⁹

TABLE 128
MOVING-AVERAGE METHOD: TWELVE MONTH CENTERED MOVING AVERAGES OF
ORIGINAL DATA
CIGARETTE CONSUMPTION DATA FROM TABLE 126*
(daily averages, 000,000 omitted)

MONTH	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936
Jan.	249.7	276.8	313.5	330.0	326.1	287.8	299.7	321.8	354.3	389.7
Feb.	254.6	280.5	315.2	331.1	323.1	286.3	301.9	325.1	356.9	393.9
Mar.	260.2	282.4	317.3	330.4	321.0	285.8	304.3	327.0	357.8	400.8
Apr.	262.1	284.4	320.7	329.8	317.6	284.4	305.8	330.1	361.1	406.5
May	263.8	286.9	323.2	327.9	314.8	283.3	305.8	336.2	365.3	408.2
June	265.3	288.4	324.9	327.0	312.8	283.0	305.3	342.1	367.7	413.9
July	267.6	291.6	326.0	326.4	310.4	282.6	309.8	343.8	370.4	419.4
Aug.	270.1	295.2	326.6	325.8	307.8	282.8	315.6	343.8	373.9	423.2
Sept.	271.7	296.7	327.9	327.2	303.8	282.8	319.4	345.2	376.8	428.2
Oct.	271.8	299.9	328.4	328.0	299.4	282.7	323.1	348.3	379.8	430.9
Nov.	271.7	305.9	327.1	328.1	294.3	288.8	322.7	351.0	381.8	432.7
Dec.	273.5	310.5	327.2	328.0	290.7	297.0	319.9	351.8	384.9	434.5

* Data for last six months of 1926 and first six months of 1937 used to compute the first six and last six moving-average figures, respectively.

Each item of the original data of Table 126 is divided by the corresponding value of the centered moving average of Table 128 to obtain the per cents of Table 129. These are presumed to be above

⁷ Table 120 and pp. 567-69.

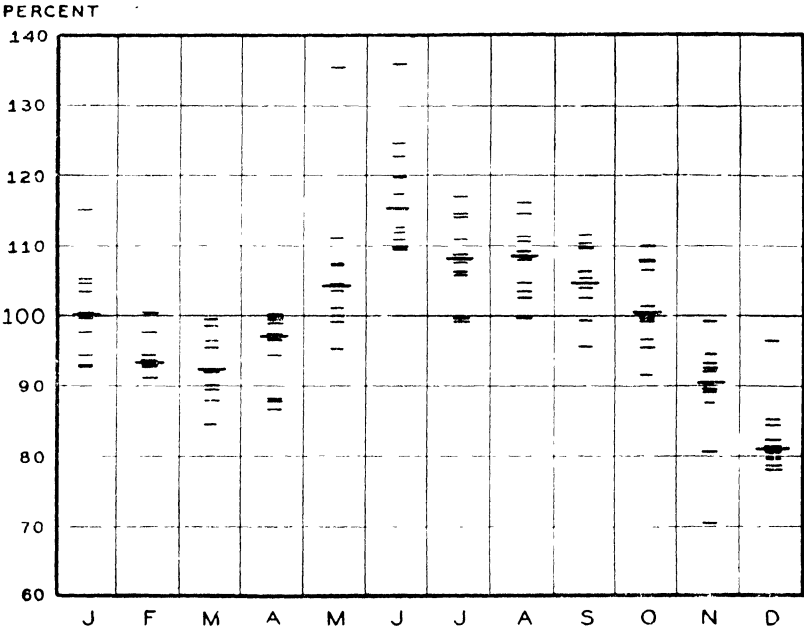
⁸ In some cases the original data must be thought of as falling at the beginning or end of the month as the case may be. But the twelve-month average will then fall at the middle of the month. In every case a moving average with an even number of months will fall midway between two monthly items of original data.

⁹ It should be noted that in this series data earlier than 1927 and more recent than the end of 1936 were available. Hence the moving average could be computed for the entire ten years. If these additional data had not been available either of two alternatives could have been followed: (1) fill in the missing six months at each end of the period free hand, (2) base the seasonal index on nine years, abandoning the first and last six months' periods.

and below 100 per cent because of the presence of seasonal variation. If the seasonal movements were completely regular in every respect, these per cents would be identical for corresponding months year after year. In practice, of course, they are not equal but, if seasonal variation in the sense previously defined is present, a seasonal pattern may be recognizable.

When the pattern is not apparent a test should be made in the same way as was done in the approximate method. The relatives of Table 129 have been arrayed in Figure 91. The grouping of the relatives is much the same as in Figure 90, indicating the similarity of the pattern obtained by the two methods. In Figure 91, however, the dashes representing the relatives for the individual months are a little better grouped than was the case in the approximate method. This result is to be expected because the relatives of the moving-average method are measured from a twelve-month moving-average base which traces the path of the combined trend and cyclical components of the series. Consequently, there is no tendency for the relatives for any month to

FIGURE 91
MOVING AVERAGE METHOD. TEST FOR SEASONAL PATTERN OF RELATIVES
OF MOVING AVERAGES



Data from Table 129

TABLE 129

MOVING-AVERAGE METHOD: ORIGINAL DATA EXPRESSED AS RELATIVES OF THE
CORRESPONDING VALUES OF THE TWELVE-MONTH MOVING AVERAGES
IN TABLE 128. SEASONAL INDEX AT THE RIGHT
CIGARETTE CONSUMPTION DATA FROM TABLE 126

MONTH	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	MEDIANS	SEASONAL INDEX (adjusted medians)
Jan. ...	93.9	97.5	104.5	99.8	92.7	100.5	92.8	115.1	103.2	105.3	100.2	100.7
Feb. ...	92.7	92.6	91.3	91.3	97.7	92.5	92.9	100.7	93.1	94.2	92.8	93.2
Mar. ...	99.3	96.7	88.3	89.5	98.5	95.3	84.5	92.1	92.0	90.1	92.1	92.5
Apr. ...	100.2	88.0	99.9	96.4	99.4	88.6	86.9	93.9	98.8	97.3	96.9	97.3
May ...	104.4	100.0	111.5	101.3	107.1	98.9	105.3	107.2	103.4	95.0	103.9	104.4
June ...	109.7	112.0	111.2	119.8	122.6	124.4	136.1	117.4	109.9	112.8	115.1	115.6
July ...	99.8	107.6	106.1	117.2	111.2	108.8	99.2	106.5	114.4	113.9	108.2	108.7
Aug. ...	111.4	116.1	108.0	104.7	99.8	109.1	114.4	110.8	103.3	102.4	108.6	109.1
Sept. ...	110.3	102.5	105.2	103.8	106.4	109.8	99.4	99.4	95.3	111.7	104.5	105.0
Oct. ...	101.5	106.7	110.0	107.7	96.5	95.3	91.6	99.3	108.0	98.8	100.4	100.9
Nov. ...	99.5	93.0	92.1	80.8	88.9	87.9	70.6	92.4	94.3	89.2	90.7	91.1
Dec. ...	81.0	78.1	81.4	85.3	80.9	79.5	78.6	84.5	82.5	96.3	81.2	81.6

scatter because of the effect of positive or negative cyclical movements. In fact the relatives for some months (notably February and December) are so closely grouped that several of the individual dashes coincide.

The seasonal pattern can be distinguished on the chart by following from month to month the long dashes which represent the position of the several medians. The pattern is also shown in Figure 94.

The numerical values for the seasonal pattern are obtained by the same final steps that were employed in the approximate method. The medians of the relatives of Table 129 are computed for each month as shown at the right of the table. The total of the medians is 1194.6. Since the final index must be adjusted to total 1200, the medians are each multiplied by the factor 1.0045.¹⁰

The resulting seasonal index shown in the last column of the table does not differ greatly from that obtained by the approximate method. Further discussion of the differences in indexes obtained by the four methods will be deferred until the other two have been explained.

Link-Relative Method¹¹

The name of this method is taken from the first step of the analysis which consists of dividing the data of each month by the data of the preceding month. The quotients expressed in per cents are above

¹⁰ As previously explained the adjustment consists of dividing each median by the average of the twelve medians. The shortest method of accomplishing this is by using the reciprocal, $1200 \div 1194.6 = 1.0045$, as a multiplier.

¹¹ This method was first published in the *Review of Economic Statistics* (Preliminary Volume I, January and April, 1919, Cambridge: Harvard University).

or below 100 per cent primarily because of the presence of seasonal movements. However, these relatives contain a small residual element of trend which, if increasing, tends to exaggerate the importance of positive seasonal movements and diminish the importance of negative seasonal movements. A decreasing trend produces the opposite effect. The relatives also contain positive and negative cyclical fluctuations which, respectively, produce effects similar to those caused by the trend. But the influence of trend and cycle in a single month is usually small compared with the seasonal change. The month-to-month relatives, therefore, express seasonal change mainly, but con-

TABLE 130

LINK-RELATIVE METHOD: ORIGINAL DATA FOR EACH MONTH EXPRESSED AS A RELATIVE OF THE PRECEDING MONTH; STEPS SHOWING COMPUTATION OF SEASONAL INDEX FROM MEDIANS OF THE LINK RELATIVES CORRECTED FOR TREND
CIGARETTE CONSUMPTION DATA FROM TABLE 126

MONTH	STEP 1									
	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936
Jan.		121.8	135.2	123.6	108.0	122.9	117.8	147.2	123.1	129.3
Feb.	100.6	96.2	87.9	91.8	104.4	91.6	100.9	88.4	90.9	90.4
Mar.	109.7	105.2	97.4	97.8	100.2	102.9	91.7	92.0	99.0	97.3
Apr.	101.5	91.7	114.3	107.5	99.8	92.5	103.3	102.9	108.4	109.5
May	104.9	114.5	112.5	104.6	106.7	111.1	155.6	116.4	105.9	98.1
June	105.7	112.6	100.3	117.9	113.8	125.6	100.4	111.4	107.0	120.4
July	91.7	97.1	95.7	97.7	90.0	87.4	74.0	91.2	104.9	102.2
Aug.	112.7	109.3	101.9	89.2	89.0	100.3	117.4	104.0	91.2	90.7
Sept.	99.6	88.7	97.8	99.6	105.2	100.6	88.0	90.1	93.0	110.4
Oct.	92.0	105.2	104.8	103.9	89.4	86.8	93.2	100.8	114.2	89.1
Nov.	97.8	88.9	83.4	75.1	90.6	94.2	77.0	93.8	87.8	90.6
Dec.	82.1	85.2	88.4	105.5	89.9	93.0	110.4	91.6	88.2	110.7

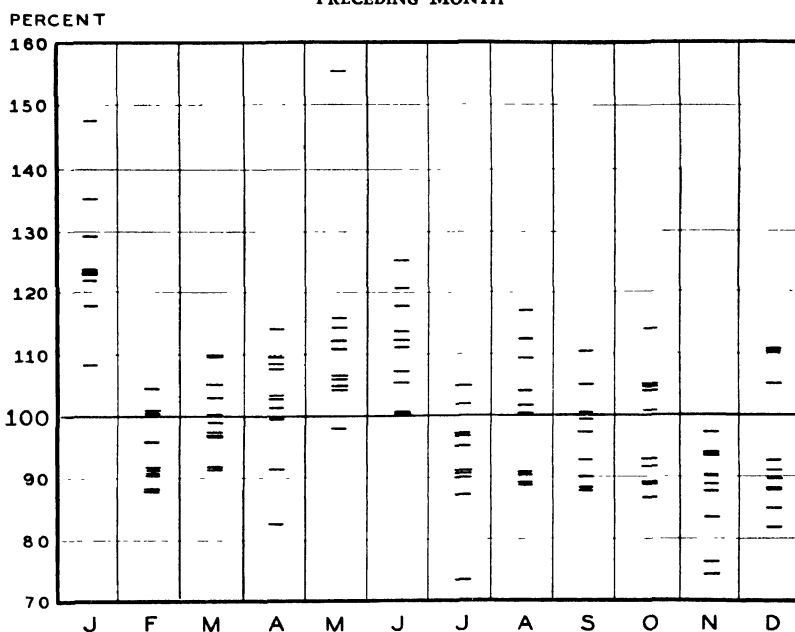
MONTH	STEP 2	STEP 3	STEP 4		STEP 5
	MEDIANS OF THE LINK RELATIVES	MEDIANS CHANGED TO FIXED JANUARY BASE	CORRECTION FOR TREND RESIDUE	CORRECTED FIXED-BASE INDEX	ADJUSTED CORRECTED FIXED-BASE INDEX
Jan.	123.1	100.0	— 0	= 100.0	102.5
Feb.	91.7	91.7	— .3	= 91.4	93.7
Mar.	98.4	90.2	— .6	= 89.7	91.9
Apr.	103.1	93.0	— .8	= 92.2	94.5
May	108.9	101.3	— 1.1	= 100.2	102.7
June	112.0	113.5	— 1.4	= 112.1	114.9
July	93.7	106.3	— 1.6	= 104.7	107.3
Aug.	101.1	107.5	— 1.9	= 105.6	108.2
Sept.	98.7	106.1	— 2.2	= 103.9	106.5
Oct.	97.0	102.9	— 2.5	= 100.4	102.9
Nov.	89.8	92.4	— 2.7	= 89.7	91.9
Dec.	90.8	83.9	— 3.0	= 80.9	82.9
Jan.	123.1	103.3	— 3.3	= 100.0	

tain a residual of trend and a residual of cycle. The subsequent steps of the method are designed to eliminate these residual elements so as to arrive at a seasonal index in the usual form. The link relatives for cigarette consumption are given in Table 130. The relative for January, 1927, is omitted to conform to the usual situation, although in this example it could have been included since the figure for December, 1926, is available.

Before proceeding with further analysis the link relatives should be tested for the existence of a seasonal pattern. This is done on the same type of graph that was used in the preceding methods. The dashes of Figure 92 are not so well concentrated as those of Figure 91 because of the presence of residues of trend and cyclical fluctuation. The test of the existence of a pattern is less simple in this case because those relatives that contain an appreciable amount of cyclical change will be far above or below the median position according to the direction of the cyclical change. The scattered dashes on the chart must be disregarded in favor of those near the median. The latter represent

FIGURE 92

LINK-RELATIVE METHOD: TEST FOR SEASONAL PATTERN OF RELATIVES OF PRECEDING MONTH



Data from Table 130.

the relation between two succeeding monthly items affected little or equally affected by cyclical change. If a majority of the dashes reflect such a situation in most of the months, a seasonal pattern is present. From this point of view June, August, and October do not show good grouping, while January, February, March, April, May, and July are well grouped. The remaining months are moderately well grouped. It should be noted that the degree of concentration of the points will indicate whether or not a seasonal component is present in the data, but since in this method each relative expresses the change from the preceding month the points of concentration of the dashes in the several months do not show the shape of the seasonal pattern.

As stated, the subsequent steps in computing the seasonal index are designed to eliminate the residues of trend and cycle present in the link relatives. The details of the several steps, as listed in Table 130, follow in order including the first step which has already been explained.

1. Divide the original data for each month by the data for the preceding month, expressing the results as per cents. This process of linking the months together is the basis for the name "link-relative method."

2. Take the median of the per cents for each month, as in the former methods. This step eliminates the effect of the cyclical fluctuation and gives a crude pattern with each month expressed as a per cent of the preceding month as a base. The median for January is computed from nine relatives since there is no relative for January, 1927.

3. Shift the chain medians to a fixed base. In practice the simplest procedure is to use January as a base. Then with January taken as 100, the February median becomes the February fixed-base index, since it is expressed on a January base. The index numbers for subsequent months are changed to the January base by direct multiplication. February cigarette consumption on the average is 91.7 per cent of that in January. March is 98.4 per cent of February; therefore it is 98.4 per cent of 91.7 per cent, or 90.2 per cent of January, hence the March index of 90.2. In the same way the fixed-base index for each month is computed. The 13th figure, 103.3, is obtained as follows: December consumption is 83.9 per cent of January, but on the average January is 123.1 per cent of December. Therefore the new January index

is 123.1 per cent of 83.9 per cent, or 103.3 per cent of the original January.

4. Correct the index for trend residue. The last computation of the preceding step expresses the new January median on the original January base. This new January index number should have a value of 100, if seasonal variation alone were present. Usually it will be either above or below 100 according to whether the series contains an increasing or decreasing trend component.¹² The variation of the new January index number from 100 is assumed to be the cumulative effect over the twelve-month period of the small amount of trend included each month in the computation of the link relatives. This residue is distributed over the twelve months in equal amounts on a cumulative basis starting with February, i.e., one-twelfth of the difference is added to or subtracted from the February relative, two-twelfths is applied to March, three-twelfths to April, and so on until twelve-twelfths is applied to the new January relative bringing it to 100.¹³

The two columns of Table 130 marked step 4 show the details of the work. The cumulative residue is 3.3;¹⁴ hence one-twelfth of this amount was subtracted from the February relative, 91.7, to give the corrected relative, 91.4. In the same way two-twelfths was subtracted to correct the March relative and so on until the total amount of the correction, 3.3, was subtracted from the new January relative of 103.3 to reduce it to 100.0, thus bringing it into agreement with the original January relative.

5. The corrected relatives must be adjusted so that they have a total of 1200. This step is equivalent to the final step taken in each of the preceding methods to center the twelve seasonal index figures so that they total 1200. The result is the seasonal index for the series. The relatives of step 4 are each multiplied by 1.025 to raise the sum from 1170.8 to 1200.

¹² If the seasonal pattern is not well defined, the medians of step 2 may not be representative of the data from which they were computed. Under these circumstances it is possible for the new January index number to be below 100 with an increasing trend, or the reverse, as users of the method have occasionally discovered. This is not a weakness of the link-relative method, but rather a warning of the weakness of trying to establish any seasonal pattern for the series in question.

¹³ In the original form of the method this correction was made according to the compound interest law instead of in equal amounts. The work is considerably more cumbersome and the additional precision does not repay the increased labor involved.

¹⁴ The calculations for this step were carried to one more decimal place than is shown in the table.

Ratio-to-Trend Method

This method unlike the preceding ones requires the removal of the trend prior to measuring the seasonal variation.¹⁵ Hence the first step is the determination of the type of trend suitable for any given series. The algebraic development of Hall's method in the original article depended upon the use of a straight line, but the explanation given here is modified so as to remove the straight-line trend requirement. In locating the trend and selecting the method of measuring it the procedure explained in the preceding chapter should be employed.

The several steps can be followed most readily by referring to the cigarette consumption data.

1. Fit a trend to the data and write the trend values for each month. A straight line has been selected as a proper description of the trend of cigarette consumption during the ten-year period, 1927-36. The computation of the trend increment is shown in Table 131. Time can be saved by computing the annual increment in the daily average data and taking one-twelfth of the annual increment as the monthly increment.

The trend increment, 11.67, of Table 131 means that the average growth per year in the daily consumption of cigarettes was 11,670,000. The monthly growth, therefore, was 972,500 cigarettes. The value of the trend at the middle of the ten-year period, i.e., at the end of 1931, is the arithmetic average of the original data, or 323,970,000 cigarettes. Since the midpoint of 120 time periods falls midway between the 60th and 61st periods, the value of the trend is 323,970,000 midway between December, 1931, and January, 1932. Hence the trend value for December, 1931, is $323,970,000 - \frac{1}{2}(972,500) = 323,500,000$ and the trend value for January, 1932, is $323,970,000 + \frac{1}{2}(972,500) = 324,500,000$. The trend values for months earlier than December, 1931, are obtained by subtracting the trend increment successively and values for months after January, 1932, are obtained by adding the trend increment successively.¹⁶ The actual work is carried out in units

¹⁵ The ratio-to-trend method was first presented in the June, 1924, issue of the *Journal of the American Statistical Association* in two articles—"Seasonal Variation as a Relative of Secular Trend," by L. W. Hall, and "The Measurement of Seasonal Variation" by Helen D. Falkner. The principles developed in the two articles are equivalent, but the computation by Hall's method is simpler as modified here.

¹⁶ Theoretically the result obtained by the short-cut process is not identical with that obtained by fitting the trend directly to the monthly data, but the error involved in the approximate computation is negligible in practice. In this case the direct computation gives an increment of .989. The maximum effect of the approximation would appear in the end values of the monthly trend in Table 132, and would amount to 59.5 (.989 —

TABLE 131

RATIO-TO-TREND METHOD: COMPUTATION OF MONTHLY INCREMENT OF STRAIGHT-LINE TREND, USING ANNUAL AVERAGES OF THE ORIGINAL MONTHLY DATA
CIGARETTE CONSUMPTION DATA FROM TABLE 126
(daily averages, 000,000 omitted)

YEAR	ΣY	Σv	ΣvY	
1927	266.2	-9	-2395.8	$\Sigma x^2 = \frac{N^3 - N}{12} = \frac{1000 - 10}{12} = 82.5$
1928	289.4	-7	-2025.8	
1929	326.1	-5	-1630.5	
1930	327.7	-3	-983.1	
1931	310.8	-1	-310.8	Trend increment = $\frac{\Sigma(vY)}{\Sigma x^2} = \frac{962.8}{82.5} = 11.67$
1932	283.0	1	283.0	
1933 ...	306.2	3	918.6	Annual increment = 11.67
1934 .	344.2	5	1721.0	
1935 .	368.8	7	2581.6	Monthly increment = $\frac{11.67}{12} = .9725$
1936	418.6	9	3767.4	
			9271.6	
			-7346.0	
			2) 1925.6	
			962.8	

of one million cigarettes as shown in Table 132. A calculating machine is a great assistance in the computation.

2. Remove the trend by dividing each item of original data by the corresponding value of the trend ($Y \div Y_t$). The results are given

TABLE 132

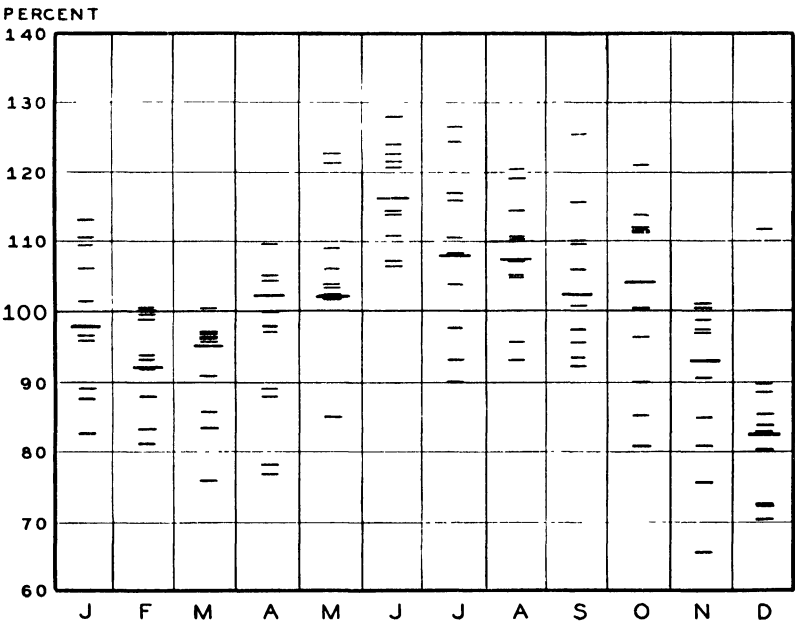
RATIO-TO-TREND METHOD: STRAIGHT-LINE TREND FITTED TO THE ORIGINAL MONTHLY DATA; MONTHLY INCREMENT = .9725, FROM TABLE 131
CIGARETTE CONSUMPTION DATA FROM TABLE 126
(daily averages, 000,000 omitted)

MONTH	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936
Jan.	266.1	277.8	289.4	301.1	312.8	324.5	336.1	347.8	359.5	371.1
Feb.	267.1	278.7	290.4	302.1	313.8	325.4	337.1	348.8	360.4	372.1
Mar.	268.1	279.7	291.4	303.1	314.7	326.4	338.1	349.7	361.4	373.1
Apr.	269.0	280.7	292.4	304.0	315.7	327.4	339.0	350.7	362.4	374.1
May	270.0	281.7	293.3	305.0	316.7	328.3	340.0	351.7	363.4	375.0
June ..	271.0	282.6	294.3	306.0	317.6	329.3	341.0	352.7	364.3	376.0
July	271.9	283.6	295.3	307.0	318.6	330.3	342.0	353.6	365.3	377.0
Aug.	272.9	284.6	296.3	307.9	319.6	331.3	342.9	354.6	366.3	377.9
Sept.	273.9	285.6	297.2	308.9	320.6	332.2	343.9	355.6	367.2	378.9
Oct.	274.9	286.5	298.2	309.9	321.5	333.2	344.9	356.5	368.2	379.9
Nov.	275.8	287.5	299.2	310.8	322.5	334.2	345.9	357.5	369.2	380.9
Dec.	276.8	288.5	300.1	311.8	323.5	335.2	346.8	358.5	370.2	381.8

.9725) = .98175 or 1.0. Hence the correct value of the trend for January, 1927, is 265.1 and for December, 1936, is 382.8. The error diminishes toward the center of the period from either extreme, but even for the end months is not important enough to justify the increased labor of the direct computation.

FIGURE 93

RATIO-TO-TREND METHOD: TEST FOR SEASONAL PATTERN OF RELATIVES OF TREND



Data from Table 133.

TABLE 133

RATIO-TO-TREND METHOD: ORIGINAL DATA EXPRESSED AS RELATIVES OF TREND;
RELATIVES CONTAIN SEASONAL AND CYCLICAL COMPONENTS;
SEASONAL INDEX AT THE RIGHT

CIGARETTE CONSUMPTION DATA FROM TABLE 126

MONTH	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	MEDIANS	SEASONAL INDEX (adjusted medians)
Jan. . .	88.1	97.2	113.2	109.4	96.6	89.1	82.7	106.5	101.7	110.6	99.5	98.3
Feb. . .	88.4	93.1	99.1	100.1	100.6	81.4	83.2	93.9	92.2	99.8	93.5	92.4
Mar. . .	96.6	97.7	96.2	97.5	100.5	83.5	76.1	86.1	91.0	96.8	96.4	95.3
Apr. . .	97.7	89.2	109.5	104.5	100.0	77.0	78.4	88.3	98.4	105.7	98.1	96.9
May . .	102.0	101.8	122.8	109.0	106.4	85.3	121.6	102.5	103.9	103.4	103.7	102.5
June . .	107.5	114.3	122.8	128.0	120.8	106.9	121.8	113.8	110.9	124.2	117.6	116.2
July . .	98.2	110.6	117.1	124.6	108.3	93.1	89.9	103.6	116.0	126.7	109.5	108.2
Aug. . .	110.3	120.4	119.0	110.8	96.1	93.1	105.2	107.4	105.5	114.6	108.9	107.6
Sept. . .	109.5	106.5	116.1	110.0	100.8	93.4	92.4	96.5	97.8	126.2	103.7	102.5
Oct. . .	100.4	111.7	121.2	113.9	89.9	80.9	85.8	96.9	111.4	112.1	105.9	104.6
Nov. . .	97.8	99.0	100.7	85.3	81.1	75.9	65.9	90.7	97.5	101.3	94.1	93.0
Dec. . .	80.1	84.0	88.8	89.7	72.7	70.4	72.5	82.9	85.8	111.9	83.5	82.5

in Table 133. The advantage of removing the trend by direct division is explained in detail in chapter XXII, page 583.

3. The relatives, $Y \div Y_{\mu}$, should be tested for the existence of a seasonal pattern, unless a pattern is evident by inspection.

The relatives are presented in the usual form in Figure 93. The relatives contain the seasonal and cyclical components, hence the position of the dashes for any month will be scattered to the extent that the presence of positive or negative cyclical fluctuation tends to move individual relatives above or below the values they would have as the result of seasonal movement. If a pattern exists, however, there should be a tendency of the dashes to concentrate somewhere between the extreme values. The dashes are well concentrated in February, March, May, August, November, and December; moderately well concentrated in April, June, and September; and poorly concentrated in January, July, and October.

4. Take the median of the relatives, $Y \div Y_{\mu}$, for each month as in the other three methods.

5. Adjust the values of the medians so that they total 1200.

This step is the equivalent of the final step of each of the other methods. The adjusted medians of the last column of Table 133 are obtained by applying the multiplier, .98814¹⁷ to the medians of the preceding column.

The Use of the Median

In all four methods of measuring seasonal the shift from individual relatives to a preliminary pattern was made by computing the median of the relatives for each month of the year. The median is preferable to other averages because each monthly relative is equally important in determining the position of the median. Further, extreme items such as those appearing in the pattern charts in several months count merely as one item above or below the median position, whereas in computing the arithmetic average such extremes would receive extra weight because of their high or low values. In some respects the mode would be superior to the median in determining a pattern because it places greater emphasis on the position at which the items are concentrated, neglecting extreme items. Points of concentration, however, may be absent in many months when as few as ten items are used,

¹⁷ $1200 \div 1213.9 = .98814$.

hence the mode is not practical unless a long period of years is used in obtaining a seasonal index.

Several modified forms of average have been proposed by statisticians to replace the simple median in computing a seasonal index. There is considerable doubt whether any of them is superior to the median for general use although in analyzing particular series circumstances may appear which point to the use of one of the following measures instead of the simple median.

1. The arithmetic average of the middle four, five, or six relatives.
2. The arithmetic average after the highest and lowest relatives or the two highest and two lowest relatives have been discarded.
3. The arithmetic average or the median determined from whatever relatives remain after discarding all extreme items. (The average or median may depend upon different numbers of relatives in the several months.)

The Results of the Four Methods

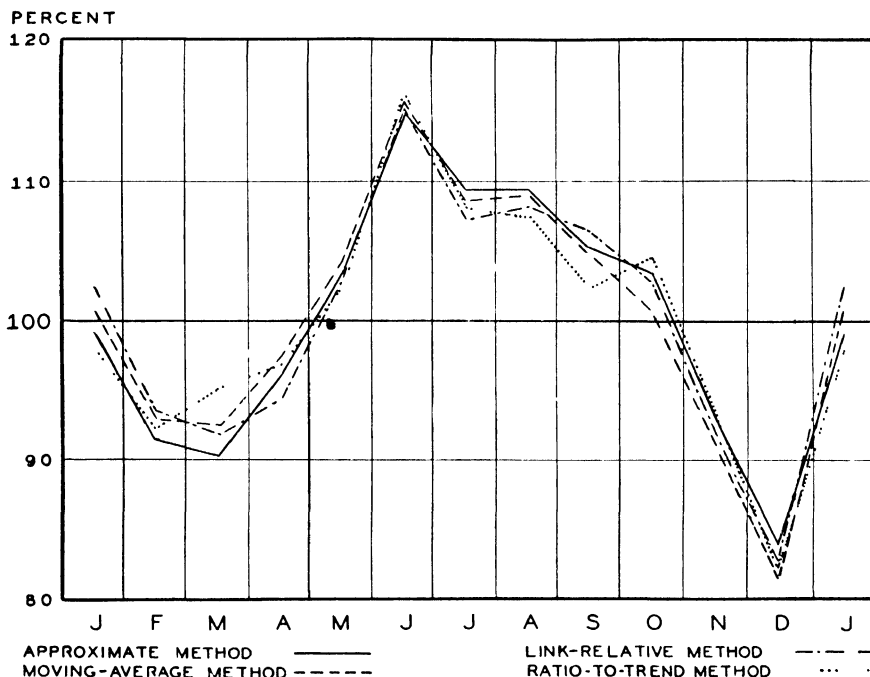
The preceding pages have been devoted exclusively to explanations of the methods of obtaining a seasonal index. An appraisal of the four methods has been deferred to this point so that the features of the four can be compared. The seasonal indexes have been reproduced in Table 134 and the four are presented in Figure 94. The pattern defined by each of the four indexes is essentially the same and in fact in some months the four curves are so close together that careful reading is required to distinguish them. Other months show appreciable differences which will be discussed presently.

TABLE 134
FINAL SEASONAL INDEXES, BY THE FOUR METHODS
CIGARETTE CONSUMPTION DATA FROM TABLE 126

MONTH	(1) APPROXIMATE METHOD	(2) MOVING- AVERAGE METHOD	(3) LINK- RELATIVE METHOD	(4) RATIO-TO- TREND METHOD
Jan.	99.3	100.7	102.5	98.3
Feb.	91.5	93.2	93.7	92.4
Mar.	90.3	92.5	91.9	95.3
Apr.	96.2	97.3	94.5	96.9
May	103.7	104.4	102.7	102.5
June	114.8	115.6	114.9	116.2
July	109.3	108.7	107.3	108.2
Aug.	109.3	109.1	108.2	107.6
Sept.	105.3	105.0	106.5	102.5
Oct.	103.4	100.9	102.9	104.6
Nov.	92.9	91.1	91.9	93.0
Dec.	84.1	81.6	82.9	82.5

FIGURE 94

SEASONAL PATTERNS OF CIGARETTE CONSUMPTION ACCORDING TO FOUR METHODS



Data from Table 134.

The pattern declines from January to an early year low in February and March followed by sharp expansion in April and May to the high point of the year in June. From this peak the curves drop sharply in July and continue a gradual contraction through August, September, and October. Sharper declines are typical during November and December, the lowest month, followed by an even sharper rise from December to January. Several variations from this general pattern are evident in the chart. The approximate method tends to be lower than the others during the early months and higher near the end of the year. This is the result of using the average for each year as a base in computing the relatives. The effect of trend for one year is left in the relatives. The ratio-to-trend method has values either higher or lower than any of the others in eight of the twelve months. This variability is the result of removing the trend before computing the seasonal index, thus increasing the effect of the cyclical fluctuations in the relatives, $Y \div Y_t$, and producing greater spread of the dashes

in Figure 93. As a result the medians are not as precisely defined as in the other methods.

Since the four seasonal indexes are available for a single series, it will be worthwhile raising the question of relative superiority. A test was made as follows: A standard for comparison was set up by taking the average of the four indexes of Table 134 for each month and subtracting the monthly values of each of the four indexes from the corresponding values of the standard. The average of the twelve differences taken without regard to sign is the average deviation dispersion of each index from the standard. The results turned out to be:

METHOD	AVERAGE DISPERSION FROM STANDARD
Approximate	.81
Moving-average	.72
Link-relative	.88
Ratio-to-trend	1.14

According to this test the moving-average method gives the best seasonal index, followed in order by the approximate, link-relative, and ratio-to-trend methods. This test must not be interpreted to mean that the moving-average method will give the best results for all series. There is, however, this much to be said in favor of the moving-average method: the cigarette consumption series has no unusual features and many series contain similar seasonal movements. Two further observations are necessary concerning choice of method of measuring seasonal variation: (1) the use of any one of the four methods to measure the seasonal of cigarette consumption would produce satisfactory results, (2) in general the choice of method can be left to the predilection of the computer, conditioned by any peculiarities present in a particular series and by the criteria discussed in the next two paragraphs.

From the standpoint of simplicity the approximate method, of course, comes first, followed by the moving-average, the ratio-to-trend, and the link-relative. The concept of the moving average is simple in itself, and its use here involves nothing more than defining the path followed by a twelve-month moving average as a description of trend and cycle combined. The ratio-to-trend method is little more involved since it differs from the moving-average method at the outset only in using trend as a base for ratios instead of trend and cycle combined. In either method the same averaging process gives the seasonal pattern. The link-relative method is somewhat more involved than the other three.

From the standpoint of amount of computation the approximate method again takes first place. The ratio-to-trend method comes next provided the purpose is to obtain the cyclical fluctuations. If the seasonal variations only are required, the computation of trend represents an extra step as compared with the link-relative. The most tedious of all to compute is the moving-average since it involves two series of additions in contrast to either one or two series of divisions in the other methods.

THE CYCLICAL REMAINDERS

The further steps necessary to obtain the cyclical fluctuations of a series depend upon the method employed in measuring the seasonal pattern. The approximate, moving-average and link-relative methods determine the seasonal pattern prior to measuring the trend, whereas the ratio-to-trend method measures and removes the trend prior to computing the seasonal pattern. So far as the ultimate objective of obtaining the cycles is concerned the order of removing these components does not matter. The actual operations involve two divisions and it becomes merely a question of the order in which they shall be performed. In symbols the steps are:

approximate method	$(Y \div Y_s) \div Y_t = Y_o$
moving-average method	$(Y \div Y_s) \div Y_t = Y_o$
link-relative method	$(Y \div Y_s) \div Y_t = Y_o$
ratio-to-trend method	$(Y \div Y_t) \div Y_s = Y_o$

in which,

Y = the original data

Y_s = the seasonal pattern (in index form, 100 base)

Y_t = the trend component

Y_o = the cyclical component (in index form, 100 base)

The computation of the cyclical fluctuations of cigarette consumption by the link-relative method is shown in Table 135. Column 1 contains the original daily average cigarette consumption copied from Table 126. Column 2 is the seasonal index copied from either Table 130 or Table 134. The index has identical values each year; hence it need not be rewritten from year to year. The effect of seasonal variation as measured by the regular pattern is removed from the series by dividing each item of original data by the corresponding seasonal index figure; thus for January, 1927, $234.5 \div 1.025 = 228.8$. The same computation for each month gives the seasonally adjusted series of column 3.

The next step in the analysis is to fit a properly selected trend to the seasonally corrected data of column 3. If a straight line is to be fitted, the work can be abbreviated by finding the annual increment and dividing it by 12 to obtain the monthly increment as was explained on page 608. The actual computation then would consist of fitting a straight line to the annual averages of column 3. But notice that these averages are practically identical with the annual averages of the original data in column 1. This relation will always hold because the seasonal index was adjusted to 1200 in order to remove a balanced seasonal curve. Therefore the trend may be fitted to the original data of column 1 or to the seasonally corrected data of column 3.¹⁸ Hence column 4 is a copy of Table 132.

The final step of the analysis consists in dividing the data of column 3, containing trend and cycle, by the trend of column 4 to obtain a set of per cents expressing the cyclical fluctuations as relatives of the trend. We divide at this point to obtain relative cycles rather than subtract to obtain absolute cycles for the same reason that we express seasonal in relative form, namely, to eliminate the tendency of the amount of cyclical fluctuation to increase with an increasing trend or to decrease with a decreasing trend.¹⁹ The computation for January, 1927, is $(228.8 \div 266.1) = .860$ or 86.0 per cent, which means that the position of the cycle at that time was 14 per cent below the average position represented by the trend. Similar computations for each month give the relative cyclical values of column 5.

The relative cycles are plotted in Figure 95. There is considerable irregularity from month to month, but the path of the cycles is fairly clear. From a point below the average level early in 1927 consumption expanded throughout 1927, 1928, and 1929. A gradual decline set in after October, 1929, and continued irregularly until the middle of 1931 when consumption had dropped to about its average level. A sharp decline during the second half of 1931 was continued through 1932. Early months of 1933 showed no improvement but the subsequent months of the year require special explanation, because the changes that occurred come under the head of irregularities. The

¹⁸ The monthly trend increment obtained by fitting to the average of column 3 and dividing by 12 turns out to be .9748, whereas the increment obtained in Table 131 is .9725. This difference is negligible in practice.

¹⁹ If a graph of a particular series shows that the cycles have no tendency toward higher amplitudes as the trend of the series increases, cyclical differences instead of relative cycles would be preferable. There is considerable question, however, whether this situation exists in practice; hence relative cycles are in accord with business experience.

TABLE 135
COMPLETE ANALYSIS OF CIGARETTE CONSUMPTION
DATA TO OBTAIN CYCLICAL COMPONENT

YEARS AND MONTHS	(1) DAILY AVERAGE CIGARETTE CONSUMPTION (000,000 omitted) Y	(2) SEASONAL INDEX LINE- RELATIVE METHOD Y_s	(3) SEASONALLY CORRECTED DATA $(1) \div (2)$ $Y \div Y_s = Y_t + e$	(4) STRAIGHT LINE TREND Y_t	(5) RELATIVE CYCLES (per cents) $(3) \div (4)$ $Y_t + e \div Y_t = Y_c$
1927 Jan.	234.5	102.5	228.8	266.1	86.0
Feb. ...	236.0	93.7	251.9	267.1	94.3
Mar. ...	258.9	91.9	281.7	268.1	105.1
Apr. ...	262.7	94.5	278.0	269.0	103.3
May ...	275.5	102.7	268.3	270.0	99.4
June ...	291.2	114.9	253.4	271.0	93.5
July ...	267.0	107.3	248.8	271.9	91.5
Aug. ...	300.9	108.2	278.1	272.9	101.9
Sept. ...	299.8	106.5	281.5	273.9	102.8
Oct.	275.9	102.9	268.1	274.9	97.5
Nov. ...	269.8	91.9	293.6	275.8	106.5
Dec.	221.6	82.9	267.3	276.8	96.6
Av.	266.2		266.6		
1928 Jan.	270.0		263.4	277.8	94.8
Feb. ...	259.7		277.2	278.8	99.4
Mar. ...	273.2		297.3	279.7	106.3
Apr. ...	250.4		265.0	280.7	94.4
May ...	286.8		279.3	281.7	99.1
June ...	323.0		281.1	282.6	99.5
July ...	313.7		292.4	283.6	103.1
Aug. ...	342.8		316.8	284.6	111.3
Sept. ...	304.2		285.6	285.6	100.0
Oct.	320.1		311.1	286.5	108.6
Nov. ...	284.5		309.6	287.5	107.7
Dec.	242.4		292.4	288.5	101.4
Av.	289.4		289.3		
1929 Jan.	327.7		319.7	289.4	110.5
Feb.	287.9		307.3	290.4	105.8
Mar.	280.3		305.0	291.4	104.7
Apr. ...	320.3		338.9	292.4	115.9
May ...	360.3		350.8	293.3	119.6
June ...	361.3		314.4	294.3	106.8
July ...	345.9		322.4	295.3	109.2
Aug. ...	352.6		325.9	296.3	110.0
Sept. ...	345.0		323.9	297.2	109.0
Oct.	361.4		351.2	298.2	117.8
Nov. ...	301.4		328.0	299.2	109.6
Dec.	266.5		321.5	300.1	107.1
Av.	326.1		325.8		
1930 Jan.	329.3		321.3	301.1	106.7
Feb.	302.3		322.6	302.1	106.8
Mar. ...	295.6		321.7	303.1	106.1
Apr. ...	317.8		336.3	304.0	110.6
May ..	332.3		323.6	305.0	106.1
June ..	391.7		340.9	306.0	111.4
July ..	382.5		356.5	307.0	116.1
Aug. ...	341.2		315.3	307.9	102.4
Sept. ...	339.7		319.0	308.9	103.3
Oct. ...	353.1		343.1	309.9	110.7
Nov. ...	265.1		288.5	310.8	92.8
Dec.	279.8		337.5	311.8	108.2
Av.	327.7		327.2		
1931 Jan.	302.2		294.8	312.8	94.2
Feb.	315.6		336.8	313.8	107.3
Mar. ...	316.2		344.1	314.7	109.3

TABLE 135—Continued
COMPLETE ANALYSIS OF CIGARETTE CONSUMPTION
DATA TO OBTAIN CYCLICAL COMPONENT

YEARS AND MONTHS	(1) DAILY AVERAGE CIGARETTE CONSUMPTION (000,000 omitted) Y	(2) SEASONAL INDEX LINK- RELATIVE METHOD Y_s	(3) SEASONALLY CORRECTED DATA $(1) \div (2)$ $Y \div Y_s = Y_t + e$	(4) STRAIGHT LINE TREND Y_t	(5) RELATIVE CYCLES (per cents) $(3) \div (4)$ $Y_t + e \div Y_t = Y_e$
Apr. ...	315.7		334.1	315.7	105.8
May	337.0		328.1	316.7	103.6
June ...	383.6		333.9	317.6	105.1
July ...	345.2		321.7	318.6	101.0
Aug.	307.1		283.8	319.6	88.8
Sept. ...	323.2		303.5	320.6	94.7
Oct.	288.9		280.8	321.5	87.3
Nov. ...	261.7		284.8	322.5	88.3
Dec.	235.3		283.8	323.5	87.7
Av.	310.8		310.8		
1932 Jan.	289.1		282.0	324.5	86.9
Feb.	264.8		282.6	325.4	86.8
Mar.	272.5		296.5	326.4	90.8
Apr.	252.1		266.8	327.4	81.5
May ...	280.2		272.8	328.3	83.1
June ...	352.0		306.4	329.3	93.0
July ...	307.5		286.6	330.3	86.8
Aug. ...	308.4		285.0	331.3	86.0
Sept. ...	310.4		291.5	332.2	87.7
Oct.	269.4		261.8	333.2	78.6
Nov.	253.8		276.2	334.2	82.6
Dec.	236.1		284.8	335.2	85.0
Av.	283.0		282.8		
1933 Jan.	278.1		271.3	336.1	80.7
Feb.	280.5		299.4	337.1	88.8
Mar.	257.2		279.9	338.1	82.8
Apr.	265.8		281.3	339.0	83.0
May ...	413.6		402.7	340.0	118.4
June ...	415.4		361.5	341.0	106.0
July ...	307.3		286.4	342.0	83.7
Aug.	360.9		333.5	342.9	97.3
Sept. ...	317.6		298.2	343.9	86.7
Oct.	296.0		287.7	344.9	83.4
Nov.	227.8		247.9	345.9	71.7
Dec.	251.6		303.5	346.8	87.5
Av.	306.2		304.4		
1934 Jan.	370.4		361.4	347.8	103.9
Feb.	327.4		349.4	348.8	100.2
Mar.	301.1		327.6	349.7	93.7
Apr.	309.8		327.8	350.7	93.5
May ...	360.5		351.0	351.7	99.8
June ...	401.5		349.4	352.7	99.1
July ...	366.3		341.4	353.6	96.5
Aug.	381.0		352.1	354.6	99.3
Sept. ...	343.1		322.2	355.6	90.6
Oct.	345.7		336.0	356.6	94.2
Nov.	324.2		352.8	357.5	98.7
Dec.	297.1		358.4	358.5	100.0
Av.	344.2		344.1		
1935 Jan.	365.7		356.8	359.5	99.2
Feb.	332.4		354.7	360.4	98.4
Mar.	329.0		358.0	361.4	99.1
Apr.	356.6		377.4	362.4	104.1
May	377.7		367.8	363.4	101.2
June ...	404.0		351.6	364.3	96.5

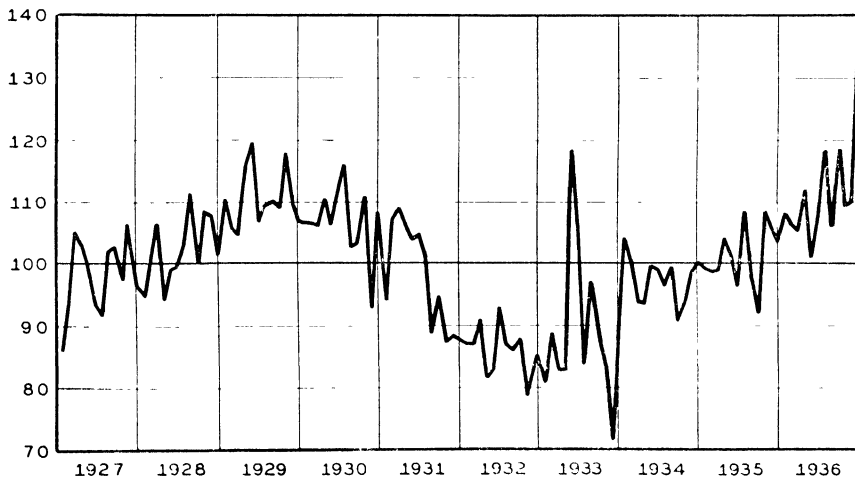
TABLE 135—Continued
COMPLETE ANALYSIS OF CIGARETTE CONSUMPTION
DATA TO OBTAIN CYCLICAL COMPONENT

YEARS AND MONTHS	(1) DAILY AVERAGE CIGARETTE CONSUMPTION (000,000 omitted) Y	(2) SEASONAL INDEX LINK- RELATIVE METHOD Y_s	(3) SEASONALLY CORRECTED DATA $(1) \div (2)$ $Y \div Y_s = Y_t + \bullet$	(4) STRAIGHT LINE TREND Y_t	(5) RELATIVE CYCLES (per cents) $(3) \div (4)$ $Y_t + \bullet \div Y_t = Y_s$
July ...	423.8		395.0	365.3	108.1
Aug. ...	386.3		357.0	366.3	97.5
Sept. ...	359.1		337.2	367.2	91.8
Oct.	410.0		398.4	368.2	108.2
Nov.	360.0		391.7	369.2	106.1
Dec.	317.5		383.0	370.2	103.5
Av.	368.8		369.0		
1936 Jan.	410.5		400.5	371.1	107.9
Feb.	371.2		396.2	372.1	106.5
Mar.	361.1		392.9	373.1	105.3
Apr.	395.6		418.6	374.1	111.9
May	387.9		377.7	375.0	100.7
June	467.0		406.4	376.0	108.1
July	477.5		445.0	377.0	118.0
Aug.	433.2		400.4	377.9	106.0
Sept.	478.1		448.9	378.9	118.5
Oct.	425.9		413.9	379.9	108.9
Nov.	385.9		419.9	380.9	110.2
Dec.	427.3		515.4	381.8	135.0
Av.	418.6		419.6		

FIGURE 95

RELATIVE CYCLES OF CIGARETTE CONSUMPTION, MONTHLY, 1927-36

PERCENT



Data from Table 135.

extreme expansion and contraction shown on the chart should not be taken as a cyclical rhythm. In anticipation of the effect of the NIRA, dealers stocked cigarettes heavily in May and June. From then until the end of the year withdrawals were small, but the extent of the decline in consumption had apparently been overestimated, hence the relatively large withdrawals in January, 1934. After that the curve resumes its gradual expansion through 1934, 1935, and 1936. The curve is definitely above its average level after September, 1935, and reaches its highest cyclical level in December, 1936.

The minor fluctuations which occur from month to month are not to be interpreted as variations in the actual consumption of cigarettes. Although these data are known as cigarette consumption they represent tax-paid withdrawals from manufacturers' warehouses by wholesalers, jobbers, etc. Hence the irregularities from month to month are an evidence of changes in stocks of cigarettes in the hands of dealers rather than changes in number of cigarettes smoked. But since dealers' stocks seldom amount to more than a few months' supply, the difference between changes in stock and changes in consumption is too slight to have any adverse effect on the use of the series as an indication of consumption in the cyclical sense.

The use of either the approximate or the moving-average method of measuring seasonal would require no change in the form of the computation for measuring cycles as shown in Table 135. The only difference would arise from the substitution of either of the other seasonal indexes for the link-relative index. The effect on the cyclical remainders would be slight, since, as has been shown in Figure 94, the three indexes do not vary greatly from each other. The ratio-to-trend method reverses the order of removing the trend and seasonal components, but otherwise parallels the computation in Table 135. The cyclical fluctuations would differ appreciably in certain months from those shown in Figure 95, because the seasonal pattern obtained by the ratio-to-trend method departs somewhat from the others.

The description and interpretation of cyclical fluctuations is the most involved part of time-series analysis. If the cyclical component possessed any regularity of amplitude or period, the problem of recognizing cyclical waves would be greatly simplified. As the matter stands statisticians must draw liberally upon their knowledge of the internal and external circumstances surrounding a series in order to distinguish cyclical movements from irregularities and minor variations present

in the data. The best guide is familiarity with the methods of computing cycles, a sound knowledge of business progress and a liberal use of common sense.

SPECIAL SEASONAL PROBLEMS

There are two problems to be discussed in relation to the concept of establishing a seasonal pattern. The first of these is the question of what to do about the seasonal pattern when a series is being subjected to continuous analysis. The second is the question of what to do about the seasonal pattern when the seasonal component of a series changes amplitude. These two problems will be discussed under the titles, moving seasonal index and changing seasonal pattern.

Moving Seasonal Index

The problem involved in continuous analysis can be explained by referring to the cigarette consumption series. The analysis has been made for the ten-year period, 1927-36. Both the trend and seasonal components have been measured for the ten years. If the analysis were to be continued into subsequent years, what should be done about these two components? Two plans immediately suggest themselves: (1) use the seasonal pattern already determined and simply project the straight-line trend, (2) recompute both the trend and the seasonal pattern as each additional year of data becomes available. Neither of these methods is satisfactory for practical purposes. The former is particularly objectionable because the projected trend may not fit the data at all. The latter has no theoretical objections but involves an impossible amount of labor. The problem of projecting the trend can be met by using a moving trend as described in the next chapter. A parallel procedure which can be used for seasonal will be explained at this point.

The seasonal pattern for cigarette consumption determined for the years, 1927-36, would be used during the year 1937. When data for the full year 1937 became available, a new seasonal pattern based on the eleven years, 1927-37, would be computed. This new pattern would be used during 1938. In the same way a pattern based on twelve years of data would be used in 1939 and so on. This method permits the seasonal component to change gradually over a period of years, hence the name moving seasonal. The change is accomplished through the shift in the position of the monthly medians

brought about by including the additional year. By referring to Figure 90, 91, 92, or 93 it should be clear that when the relatives of the individual months are closely grouped about their medians the addition of one relative, regardless of the value, would have comparatively little effect upon the median and therefore upon the seasonal pattern. It is quite possible that the addition of two years of data would leave unchanged all of the medians and accordingly the seasonal pattern. For that reason there is considerable justification for moving the seasonal pattern up every two years instead of every year.

The effect of changing the number of years used in computing the seasonal index of cigarette consumption can be studied in Table 136. Column 1 is a reproduction from Table 130 of the seasonal

TABLE 136

SEASONAL INDEXES OF CIGARETTE CONSUMPTION BY THE LINK-RELATIVE METHOD BASED ON DIFFERENT TIME PERIODS

MONTH	(1) TEN YEARS 1927-36	(2) ELEVEN YEARS 1927-37	(3) TWELVE YEARS 1927-38	(4) TEN YEARS 1928-37	(5) TEN YEARS 1929-38
Jan.	102.5	102.9	100.8	103.1	102.3
Feb.	93.7	94.5	94.7	94.1	94.1
March	91.9	92.4	93.0	91.2	92.0
April	94.5	95.1	95.0	93.5	95.1
May	102.7	101.4	103.2	101.4	103.5
June	114.9	113.1	115.5	114.1	116.0
July	107.3	108.2	108.1	109.8	108.8
Aug.	108.2	108.6	109.3	108.5	108.5
Sept.	106.5	108.1	107.7	106.5	107.3
Oct.	102.9	100.8	99.6	102.6	98.9
Nov.	91.9	91.3	90.1	91.2	90.0
Dec.	82.9	83.7	82.9	84.2	83.5

index for the years 1927-36 by the link-relative method. The index for the eleven years 1927-37 is given in column 2 and for the twelve years 1927-38 in column 3. These three indexes do not vary much in most of the months, but it is worth noting that the greatest variation occurs in the months that showed poor grouping of the relatives in Figure 92, namely, June, September, and October. According to the argument of the preceding paragraph there should be better agreement between the twelve-year index and the ten-year index than between the eleven-year index and the ten-year index. This proves not to be true for the particular years included in the example because the link relatives (omitted because of lack of space) for 1938 are very different from those of other years. The differences are the result

of sharp cyclical decline and recovery during the year. Although this example fails to substantiate the argument for moving the seasonal index up at two-year intervals, the two-year interval is preferable in general.

A variant of this method consists of dropping a year at the beginning of the period each time a new year is added at the end. With this plan the seasonal pattern would always be based on the same number of years instead of being based on an increasing number of years as explained in preceding paragraphs. Thus the seasonal pattern of cigarette consumption to be used in 1938 might be based on the years 1928–37 instead of the years 1927–37, and the index for 1939 on the years 1929–38. These two indexes are presented in columns 4 and 5 of Table 136. They show the same general agreement of pattern with the index of column 1 as was found in the increasing base indexes of columns 2 and 3. The ten-year indexes also exhibit in certain months instability similar to that found in the increasing base indexes. Comparison of column 2 with 4, and 3 with 5 leads to the conclusion that for this series of data there is not much to choose between the index with an increasing number of years used in the computation and one with a fixed number of years.

The question of which to use in a particular case may be determined by the number of years of data available or the nature of the seasonal movements present in the data. Another factor of importance is the disturbed condition of the business structure throughout the decade 1930–40. When it is clear, as may be seen in the chart of sales of F. W. Woolworth Co., Figure 75, page 546, that a well-established seasonal movement is temporarily distorted by unusual conditions, the proper plan is to include as many predepression years as are available in establishing a pattern. On the other hand predepression years should be omitted in obtaining a pattern for series that exhibit permanently altered seasonal movements as a result of business conditions during the decade 1930–40.

Changing Seasonal Pattern

Sometimes external factors produce a change in the seasonal pattern of a series. Examples of such factors are new inventions, improved processes of production, planned change in monthly production schedules, change in habits of consumers, and new laws. When the change in seasonal pattern brought about by the introduction of one or more

of these factors works itself out gradually over a period of years, the moving seasonal pattern described in the preceding section will adjust itself to the change automatically. Those cases which result in more abrupt changes in the seasonal pattern require special methods of analysis. Beginning in 1934 the introduction of new models in the automobile industry was changed from January to October and November. This shift caused a marked change in the time of occurrence of the seasonal patterns of both production and sales. Similarly, quick freezing of perishable fruits and vegetables has changed the shape of the seasonal curve of marketing of these commodities.

There are two methods of dealing with such changes. If the change is abrupt, such as that which occurred in the automobile industry, the seasonal pattern should be determined separately for the years since the change. In other cases in which the change is progressive during a period of several years or is continuously progressive, special methods should be used. Several such methods²⁰ have been developed but a detailed explanation of their computation and use would carry us somewhat beyond the scope of this book. References to the original sources have been included in the list at the end of the chapter.

PROBLEMS

1. Draw rough sketches of series containing, (a) seasonal variation but no seasonal pattern, (b) a seasonal pattern, (c) no seasonal variation.
2. The following hypothetical series of quarterly data is intended for classroom use in explaining the methods of measuring seasonal variation. Compute the seasonal pattern by each of the four methods presented in the chapter.

QUARTER	1ST YEAR	2D YEAR	3D YEAR	4TH YEAR
1st	22	24	25	28
2d	13	14	16	17
3d	1	3	4	7
4th	16	17	19	22

²⁰ The principle employed may be stated as follows: relatives are obtained the same as in methods previously explained, but instead of establishing a seasonal pattern, a pattern of change is established from the relatives of each month for all years of the series. Thus a trend line fitted to the January relatives would provide a measure of the changing January seasonal for the period and similarly for the other months. The values of these twelve trend lines for any year would give the crude seasonal index for that year. Adjustment of the crude index gives the final seasonal index for the year. Continuation of this process leads to the changing seasonal index for the whole series.

3. The following is an index of hardware sales by about 150 wholesalers in the United States (January, 1938 = 100) :

MONTH	1938	1939	1940
January	100.0	107.7	115.4
February	100.0	100.3	112.1
March	127.9	131.3	133.7
April	124.6	128.0	143.2
May	123.2	142.7	148.9
June	123.9	139.0	145.8
July	111.3	124.4	135.6
August	126.1	136.6	141.8
September	138.0	164.2	151.3
October	138.6	164.0	167.9
November	129.6	148.4	154.5
December	121.3	133.7	153.0

- a) Compute the index of seasonal variation by one of the four methods described in the text. (Instructors may elect to have all students use the same method or to assign the methods individually.)
 - b) Remove the seasonal component from the series.
 - c) Fit a suitable trend to the series.
 - d) Plot the relative cycles.
 - e) Describe the cyclical movements with emphasis on amplitude and period.
4. a) If trend is fitted to annual production data, explain how to obtain the corresponding trend increment for monthly data; for quarterly data.
- b) If trend is fitted to average monthly production each year, explain how to obtain the corresponding trend increment for monthly data.
5. Apply the testing method described on page 614 to the results of Problems 2 and/or 3 to determine which measure of seasonal is most reliable.
6. By use of the references at the end of the chapter or other sources find three methods of measuring seasonal variation other than those given in the text. Describe one of the additional methods in detail emphasizing the differences from the methods with which you are familiar.
7. In chapter XXII emphasis was placed on the difference between historical trend and trend for projection. What is the corresponding problem in dealing with seasonal variation? How should the problem be solved?

REFERENCES

- BAUMAN, A. O., "Thirteen-Months-Ratio-First-Difference Method of Measuring Seasonal Variation," *Journal of the American Statistical Association*, Vol. XXIII (September, 1928).
- CARMICHAEL, F. L., "Methods of Computing Seasonal Indexes: Constant and Progressive," *Journal of the American Statistical Association*, Vol. XXII (September, 1927).

CRUM, W. L., "Progressive Variation in Seasonality," *Journal of the American Statistical Association*, Vol. XX (March, 1925).

FALKNER, HELEN D., "The Measurement of Seasonal Variation," *Journal of the American Statistical Association*, Vol. XIX (June, 1924).

HALL, LINCOLN W., "Seasonal Variation As A Relative of Secular Trend," *Journal of the American Statistical Association*, Vol. XIX (June, 1924).

HART, WILLIAM L., "The Method of Monthly Means for the Determination of a Seasonal Variation," *Quarterly Publications of the American Statistical Association*, Vol. XVIII (September, 1922).

HOMAN, JEANETTE, "Adjusting for the Changing Date of Easter in Economic Series," *Journal of the American Statistical Association*, Vol. XXVIII (September, 1933).

JOY, ARYNESS, and THOMAS, WOODLIEF, "Use of Moving Averages in the Measurement of Seasonal Variations," *Journal of the American Statistical Association*, Vol. XVIII (September, 1928).

KING, WILLFORD I., "An Improved Method for Measuring the Seasonal Factor," *Journal of the American Statistical Association*, Vol. XIX (September, 1924).

KUZNETS, SIMON, "Seasonal Pattern and Seasonal Amplitude: Measurement of Their Short-Time Variations," *Journal of the American Statistical Association*, Vol. XXVII (March, 1932).

MITCHELL, WESLEY C., *Business Cycles, The Problem and Its Setting*. New York: National Bureau of Economic Research, 1927.

Types of seasonal variation and causes are discussed on pp. 233-249.

SPURR, WILLIAM A., "A Graphic Method of Measuring Seasonal Variation," *Journal of the American Statistical Association*, Vol. XXXII (June, 1937).

WELCH, EMMETT H., "Adjusting Indexes of Seasonal Variation for Secular Trend," *Journal of the American Statistical Association*, Vol. XXV (December, 1930).

CHAPTER XXIV

SUMMARY OF THE ANALYSIS OF TIME SERIES (AN EXAMPLE)

THE PURPOSE of this chapter is to bring all of the steps of analysis together in one place by applying them to a single series of monthly data. The series used is bank debits in 140 cities exclusive of New York City as compiled by the Board of Governors of the Federal Reserve System. The reports from 140 cities include a large percentage of the total check transactions of the entire country. New York City debits are excluded because they are heavily weighted with transactions arising from activity in the stock, commodity and produce markets.

The problem that is being solved in this chapter is to separate the components of a time series and thus as the final step to obtain the cyclical fluctuations. Since bank debits are frequently used to show the course of business in general, the final result could be thought of as a measure of the cycles of American business for the period included in the analysis. This point will be discussed in more detail at the end of the chapter.

The series starts with 1919 and is carried into 1938 to the time the computations were made. The methods employed permit the continuation of the work at any time without recomputation of the part presented here.

The order of the successive steps is: adjustment of calendar variation, adjustment of changes in the price level, adjustment of seasonal variation, adjustment of trend.

ADJUSTMENT OF CALENDAR VARIATION

The dollar value of check transactions from month to month is affected directly by the number of banking days in the month. Variability from this source is eliminated by changing the data from monthly totals to daily averages.

The first column of Table 137 contains the total debits monthly in millions of dollars from January, 1919, through June, 1938. Column 2 contains the number of banking days in each month. There

TABLE 137

SUCCESSIVE STEPS IN THE ADJUSTMENT OF BANK DEBITS FOR CALENDAR VARIATION, PRICE CHANGES, SEASONAL VARIATION, AND TREND
TO OBTAIN THE CYCLICAL FLUCTUATIONS, MONTHLY DATA JANUARY, 1919-JUNE, 1938

YEAR AND MONTH	(1) MONTHLY TOTAL (000,000 omitted)	(2) NO. OF BANK- ING DAYS	(3) DAILY AVERAGE (000,000 omitted) $(1) \div (2)$	(4) INDEX OF WHOLESALE PRICES (1926=100)	(5) CORRECTED FOR PRICE CHANGE (000,000 omitted) $Y =$ $(3) \div (4)$	(6) LINK RELA- TIVES	(7) SEASONAL INDEX FROM TABLE 138 Col. 4 Y_s	(8) CORRECTED SEASONAL TENDENCY (000,000 omitted) $Y_{st} =$ $(5) \div (7)$	(9) ANNUAL TREND FROM TABLE 139 Col. 11 (000,000 omitted)	(10) MONTHLY TENDENCY (000,000 omitted) Y_t	(11) RELATIVE CYCLES $Y_c =$ $(8) \div (10)$
1919											
January	16,810	26	646.54	134.4	481.06		102.81	467.91		501.20	93.4
February	13,671	22	621.41	129.8	478.74	99.5	102.53	466.93		505.44	92.4
March	15,472	26	595.08	131.3	453.22	94.7	98.49	460.17		509.68	90.3
April	15,363	26	590.88	133.0	444.27	98.0	98.24	452.23		513.93	88.0
May	16,697	26	642.19	135.3	474.64	106.8	98.20	483.34		518.17	93.3
June	17,641	25	705.64	135.6	520.38	109.6	100.99	515.28		522.41	98.6
July	18,629	26	716.50	141.1	507.80	97.6	99.06	512.62	524.53	526.65	97.3
August	17,668	26	679.54	144.3	470.92	92.7	89.12	528.41		530.89	99.5
September	18,086	25	723.44	141.1	512.71	108.9	96.59	530.81		535.13	99.2
October	20,248	26	778.77	141.6	549.98	107.3	103.41	531.84		539.38	98.6
November	19,185	23	834.13	144.5	577.25	105.0	102.48	563.28		543.62	103.6
December	21,705	26	834.81	150.5	554.69	96.1	108.08	513.22		547.86	93.7
1920											
January	21,731	26	835.81	157.7	530.00	95.5		515.51		552.10	93.4
February	17,734	22	806.09	157.1	513.11	96.8		500.45		556.34	90.0
March	21,146	27	783.19	158.6	493.81	96.2		501.38		560.58	89.4
April	20,324	26	781.69	165.5	472.32	95.6		480.78		564.83	85.1
May	19,676	25	787.04	167.2	470.72	99.7		479.35		569.07	84.2
June	20,541	26	790.04	166.5	474.50	100.8		469.85	575.43	573.31	82.0
July	20,805	26	800.19	165.8	482.62	101.7		487.20		577.55	84.4
August	18,904	26	727.08	161.4	450.48	93.3		505.48		581.79	86.9
September	19,779	25	791.16	155.2	509.77	113.2		527.77		586.03	90.1
October	20,891	25	835.64	144.2	579.50	113.7		560.39		590.28	94.9
November	19,525	24	813.54	133.4	609.85	105.2		595.09		594.52	100.1
December	20,540	26	790.00	120.7	654.52	107.3		605.59		598.76	101.1

TABLE 137—(Continued)
SUCCESSIVE STEPS IN THE ADJUSTMENT OF BANK DEBITS FOR CALENDAR VARIATION, PRICE CHANGES, SEASONAL VARIATION, AND TREND
TO OBTAIN THE CYCLICAL FLUCTUATIONS, MONTHLY DATA JANUARY, 1919—JUNE, 1938

YEAR AND MONTH	(1) MONTHLY TOTAL (000,000 omitted)	(2) NO. OF BANK- ING DAYS	(3) DAILY AVERAGE (000,000 omitted) (1)÷(2)	(4) INDEX OF WHOLESALE PRICES (1926=100)	(5) CORRECTED FOR PRICE CHANGE (000,000 omitted) $Y_4 = (3) \div (4)$	(6) LINK RELA- TIVES	(7) SEASONAL INDEX FROM TABLE 138 COL. 4 Y_4	(8) CORRECTED FOR SEASONAL (000,000 omitted) $Y_4 + (7)$	(9) ANNUAL TREND FROM TABLE 139 COL. 11 (000,000 omitted)	(10) MONTHLY TREND (000,000 omitted) Y_4	(11) RELATIVE CYCLES $Y_6 = (8) \div (10)$
1924											
January	19,384	26	745.54	99.6	748.53	90.1		728.07		755.70	96.3
February	17,512	23	761.39	99.7	763.68	102.0		744.84		759.94	98.0
March	19,193	26	738.19	98.5	749.43	98.1		760.92		764.18	99.6
April	18,865	26	725.58	97.3	745.71	99.5		759.07		768.43	98.8
May	18,639	26	716.88	95.9	747.53	100.2		761.23		772.67	98.5
June	18,304	25	732.16	94.9	771.51	103.2		763.95		776.91	98.3
July	18,662	26	717.77	95.6	750.81	103.2		757.93	779.03	781.15	97.0
August	17,776	26	683.69	97.0	704.84	93.9		790.89		785.39	100.7
September ..	18,238	25	729.52	97.1	751.31	106.6		777.83		789.63	98.5
October	20,912	26	804.31	98.2	819.05	109.0		792.04		793.88	99.8
November ..	18,846	23	819.39	99.1	826.83	100.9		806.82		798.12	101.1
December ...	21,830	26	839.62	101.5	827.21	100.0		765.37		802.36	95.4
1925											
January	22,300	26	857.69	102.9	833.52	100.8	102.81	810.74		806.60	100.5
February	18,593	22	845.14	104.0	812.63	97.5	102.53	792.58		810.84	97.7
March	21,241	26	816.96	104.2	784.03	96.5	98.49	796.05		815.08	97.7
April	20,613	26	792.81	101.9	778.03	99.2	98.24	791.97		819.33	96.7
May	20,417	25	816.68	101.6	803.82	103.3	98.20	818.55		823.57	99.4
June	21,701	26	834.65	103.0	810.34	100.8	100.99	802.40		827.81	96.9
July	21,579	26	829.96	104.3	795.74	98.2	99.06	803.29	829.93	832.05	96.5
August	19,869	26	764.19	103.9	735.51	92.4	89.12	825.30		836.29	98.7
September ..	20,895	25	835.80	103.4	808.32	109.9	96.59	836.86		840.53	99.6
October	24,039	26	924.58	103.6	892.45	110.4	103.41	863.02		844.78	102.2
November ..	21,358	23	928.61	104.5	888.62	99.6	102.48	867.12		849.02	102.1
December ...	24,086	26	926.38	103.4	895.92	100.8	108.08	828.94		853.26	97.1

TABLE 137—(Continued)

1926	23,607	25	944.28	103.2	915.00	102.1		889.99	857.50	103.8
January	23,607	25	944.28	103.2	915.00	102.1		889.99	857.50	103.8
February	20,102	22	913.73	102.0	895.81	97.9		873.71	861.74	101.4
March	23,458	27	868.81	100.6	863.63	96.4		876.87	865.98	101.3
April	22,537	26	866.81	100.3	864.22	100.1		879.70	870.23	101.1
May	21,449	25	857.96	100.5	853.69	98.8		869.34	874.47	99.4
June	22,466	26	864.08	100.4	860.64	100.8		852.20	878.71	97.0
July	23,300	26	896.15	99.5	900.65	104.6		909.20	882.95	103.0
August	20,778	26	799.15	99.1	806.41	89.5		904.86	887.19	102.0
September	21,336	25	853.44	99.7	856.01	106.2		886.23	891.43	99.4
October	23,780	25	931.20	99.4	956.94	111.8		925.38	895.68	103.3
November	21,593	24	899.71	98.4	914.34	95.5		892.21	899.92	99.1
December	24,493	26	942.04	97.9	962.25	105.2		890.31	904.16	98.5
1927										
January	23,456	25	938.24	96.5	972.27	101.0		945.70	908.40	104.1
February	20,781	22	944.59	95.8	986.00	101.4		961.67	912.64	105.4
March	24,026	27	889.85	94.7	939.65	95.3		954.06	916.88	104.1
April	23,576	26	906.77	94.1	963.62	102.6		980.88	921.13	106.5
May	22,873	25	914.92	94.2	971.25	100.8		989.05	925.37	106.9
June	23,812	26	915.85	94.1	973.27	100.2		963.73	929.61	103.7
July	22,932	25	917.28	94.3	972.73	99.9		981.96	933.85	105.2
August	22,048	27	816.59	95.2	857.76	88.2		962.48	938.09	102.6
September	23,381	25	935.24	96.3	971.17	113.2		1005.46	942.33	106.7
October	25,111	25	1004.44	96.6	1039.79	107.1		1005.50	946.58	106.2
November	23,803	24	991.79	96.3	1029.90	99.0		1004.98	950.82	105.1
December	26,503	26	1019.35	96.4	1057.42	102.7		978.37	955.06	102.4
1928										
January	25,001	25	1000.04	96.4	1037.39	98.1	102.81	1009.04	959.30	105.2
February	21,753	23	945.78	95.8	987.24	95.2	102.53	962.88	963.54	99.9
March	25,847	27	957.30	95.5	1002.41	101.5	98.49	1017.78	967.78	105.2
April	25,225	25	1009.00	96.6	1044.51	104.2	98.24	1063.22	972.03	109.4
May	26,346	26	1013.31	97.5	1039.29	99.5	98.20	1058.34	976.27	108.4
June	27,029	26	1039.58	96.7	1075.06	103.4	100.99	1064.52	980.51	108.6
July	23,897	25	955.88	97.4	981.40	91.3	99.06	990.71	984.75	100.6
August	23,401	27	866.70	97.6	888.01	90.5	89.12	996.42	988.99	100.8
September	24,450	24	1018.75	98.6	1033.22	116.4	96.59	1069.70	993.23	107.7
October	27,705	26	1065.58	96.7	1101.94	106.7	103.41	1065.60	997.48	106.8
November	25,880	24	1078.33	95.8	1125.61	102.1	102.48	1098.37	1001.72	109.6
December	29,659	25	1186.36	95.8	1238.37	110.0	108.08	1145.79	1005.96	113.9

TABLE 137—(Continued)

SUCCESSIVE STEPS IN THE ADJUSTMENT OF BANK DEBITS FOR CALENDAR VARIATION, PRICE CHANGES, SEASONAL VARIATION, AND TREND TO OBTAIN THE CYCLICAL FLUCTUATIONS, MONTHLY DATA JANUARY, 1919-JUNE, 1938

YEAR AND MONTH	(1) MONTHLY TOTAL (000,000 omitted)	(2) NO. OF BANKING DAYS	(3) DAILY AVERAGE (000,000 omitted) $(1) \div (2)$	(4) INDEX OF WHOLESALE PRICES (1926=100)	(5) CORRECTED FOR PRICE CHANGES (000,000 omitted) $Y = (3) \div (4)$	(6) LINK RELATIVES	(7) SEASONAL INDEX FROM TABLE 138 COL. 4 Y_s	(8) CORRECTED FOR SEASONAL TRENDS (000,000 omitted) $Y_{1+e} = (5) \div (7)$	(9) ANNUAL TREND FROM TABLE 139 COL. 11 (000,000 omitted)	(10) MONTHLY TRENDS (000,000 omitted) Y_t	(11) RELATIVE CYCLES $Y_c = (8) \div (10)$
1929											
January . . .	28,095	26	1080.58	95.9	1126.78	91.0		1095.98		1010.20	108.5
February . .	24,489	22	1113.14	95.4	1166.81	103.6		1138.02		1014.44	112.2
March	28,099	26	1080.73	96.1	1124.59	96.4		1141.83		1018.68	112.1
April	26,770	26	1029.62	95.5	1078.14	95.9		1097.46		1022.93	107.3
May	26,492	26	1018.92	94.7	1075.95	99.8		1095.67		1027.17	106.7
June	26,404	25	1056.16	95.2	1109.41	103.1		1098.53		1031.41	106.5
July	28,416	26	1092.92	96.5	1132.56	102.1		1143.31	1033.53	1035.65	110.4
August	28,310	27	1048.52	96.3	1088.81	96.1		1221.73		1039.89	117.5
September . .	27,274	24	1136.42	96.1	1182.54	108.6		1224.29		1044.13	117.3
October	32,202	26	1238.54	95.1	1302.36	110.1		1259.41		1048.38	120.1
November . . .	28,486	24	1186.92	93.5	1269.43	97.5		1238.71		1032.62	117.7
December . . .	26,902	25	1076.08	93.3	1153.35	90.9		1067.13		1036.86	101.0
1930											
January	25,691	26	988.12	92.5	1068.24	92.6		1039.04		1061.10	97.9
February . . .	21,508	22	977.64	91.4	1069.63	100.1		1043.24		1065.34	97.9
March	24,983	26	960.88	90.2	1065.28	99.6		1081.61		1069.58	101.1
April	24,315	26	935.19	90.0	1039.10	97.5		1037.72		1073.83	98.5
May	24,388	26	938.00	88.8	1056.31	101.7		1075.67		1078.07	99.8
June	24,621	25	984.84	86.8	1134.61	107.4		1123.49	1084.43	1082.31	103.8
July	23,145	26	890.19	84.4	1054.73	93.0		1064.74		1086.55	98.0
August	20,941	26	805.42	84.3	955.42	90.6		1072.06		1090.79	98.3
September . .	21,253	25	850.12	84.4	1007.25	105.4		1042.81		1095.03	95.2
October	23,679	26	910.73	83.0	1097.27	108.9		1061.09		1099.28	96.5
November . . .	19,686	23	855.91	81.3	1052.78	95.9		1027.30		1103.52	93.1
December . . .	23,107	26	888.73	79.6	1116.49	106.1		1033.02		1107.76	93.3

TABLE 137—(Continued)
SUCCESSIVE STEPS IN THE ADJUSTMENT OF BANK DEBITS FOR CALENDAR VARIATION, PRICE CHANGES, SEASONAL VARIATION, AND TREND
TO OBTAIN THE CYCLICAL FLUCTUATIONS, MONTHLY DATA JANUARY, 1919—JUNE, 1938

YEAR AND MONTH	(1) MONTHLY TOTAL (000,000 omitted)	(2) NO. OF BANK- ING DAYS	(3) DAILY AVERAGE (000,000 omitted) (1)÷(2)	(4) INDEX OF WHOLESALE PRICES (1926=100)	(5) CORRECTED FOR PRICE CHANGE (000,000 omitted) $Y=(3)÷(4)$	(6) LINK RELA- TIVES	(7) SEASONAL INDEX FROM TABLE 138 COL. 4 Y_s	(8) CORRECTED FOR SEASONAL CHANGE (000,000 omitted) $Y+=(5)-(7)$	(9) ANNUAL TREND FROM TABLE 139 COL. 11 (000,000 omitted)	(10) MONTHLY TREND (000,000 omitted) Y_t	(11) RELATIVE CYCLES $Y_c=(8)÷(10)$
1934											
January	13,198	26	507.62	72.2	703.07	93.7	102.81	683.85		951.04	71.9
February	11,784	22	535.64	73.6	727.77	103.5	102.53	709.81		945.18	75.1
March	14,077	27	521.37	73.7	707.42	97.2	98.49	718.27		939.33	76.5
April	14,278	25	571.12	73.3	779.15	110.1	98.24	793.11		933.47	85.0
May	14,105	26	542.50	73.7	736.09	94.5	98.20	749.58		927.61	80.8
June	14,754	26	567.46	74.6	760.67	103.3	100.99	753.21		921.76	81.7
July	13,910	25	556.40	74.8	743.85	97.8	99.06	750.91	918.83	916.61	81.9
August	13,421	27	497.07	76.4	650.62	87.5	89.12	730.05		912.18	80.0
September	12,888	24	537.00	77.6	692.01	106.4	96.59	716.44		907.75	78.9
October	14,465	26	556.35	76.5	727.25	105.1	103.41	703.27		903.32	77.9
November	13,409	23	583.00	76.5	762.09	104.8	102.48	743.65		898.89	82.7
December	15,701	25	628.04	76.9	816.70	107.2	108.08	755.64		894.46	84.5
1935											
January	14,983	26	576.27	78.8	731.31	89.5		711.32		890.03	79.9
February	13,111	22	595.95	79.5	749.62	102.5		731.12		885.60	82.6
March	15,754	26	605.92	79.4	763.12	101.8		774.82		881.17	87.9
April	15,645	26	601.73	80.1	731.22	98.4		764.68		876.74	87.2
May	15,557	26	598.35	80.2	746.07	99.3		759.75		872.31	87.1
June	15,808	25	632.32	79.8	792.38	106.2		784.61	865.66	867.88	90.4
July	16,550	26	636.54	79.4	801.69	101.2		809.30		864.92	93.6
August	15,536	27	575.41	80.5	714.80	89.2		802.06		863.43	92.9
September	15,016	24	625.67	80.7	775.30	108.5		802.67		861.94	93.1
October	16,844	26	647.85	80.5	804.78	103.8		778.24		860.45	90.4
November	16,685	23	725.43	80.6	900.04	111.8		878.26		858.96	102.2
December	18,676	25	747.04	80.9	923.41	102.6		854.38		857.47	99.6

TABLE 137—(Continued)

is some variation in bank holidays in different cities but the ones deducted here are standard for nearly all parts of the country. Column 3 gives the daily average debits obtained by dividing each item of column 1 by the corresponding item of column 2.

The effect of the adjustment can be seen in Figure 96-A. The solid curve represents the monthly total debits from column 1 of the table and the broken curve the daily average debits from column 3. The scale at the right of the chart was made 25 times as great as the scale at the left in order to bring the two curves together. Study of the curves reveals that the greatest adjustments occur in February, August, and November. The decline in the monthly totals from January to February disappears in the daily averages. A decline in monthly totals in November also disappears in the daily averages. An opposite situation occurs in August, the daily averages bringing out the decline that is partially concealed in the monthly totals. For other months the adjustments consist mainly in smoothing the effects of four and five week-ends.

ADJUSTMENT OF CHANGES IN THE PRICE LEVEL

If the general level of prices increases, an expanded dollar value of checks will be needed to carry on the same amount of business and vice versa. Such changes in the price level may or may not accompany corresponding movements of business activity. Therefore when bank debits are used as a measure of business activity, they should be adjusted for changes in the price level. The difficulty of selecting the proper index of price change was discussed in chapter XIX (pp. 495-97). Bank debits represent to a large extent the financial transactions related to manufacturing and marketing of goods. Consequently the Bureau of Labor Statistics Index of Wholesale Prices of Commodities affords a satisfactory measure of price change for use in the adjustment of bank debits.¹

The Bureau of Labor Statistics Index is reproduced in column 4 of Table 137. The adjustment is carried out by dividing each item of column 3 by the corresponding item of column 4 to give column 5. The effect of this adjustment is to reduce the range of fluctuation of

¹ An adjusting measure such as Carl Snyder's Index of the General Price Level might be preferable in some respects because it includes prices from a much wider range of business activities than the Bureau of Labor Statistics Index. The latter has been employed in adjusting bank debits because the entire analysis is planned for computation of the current figure by the 8th to the 10th of the succeeding month. Snyder's Index is not published quickly enough to be used for this purpose.

FIGURE 96

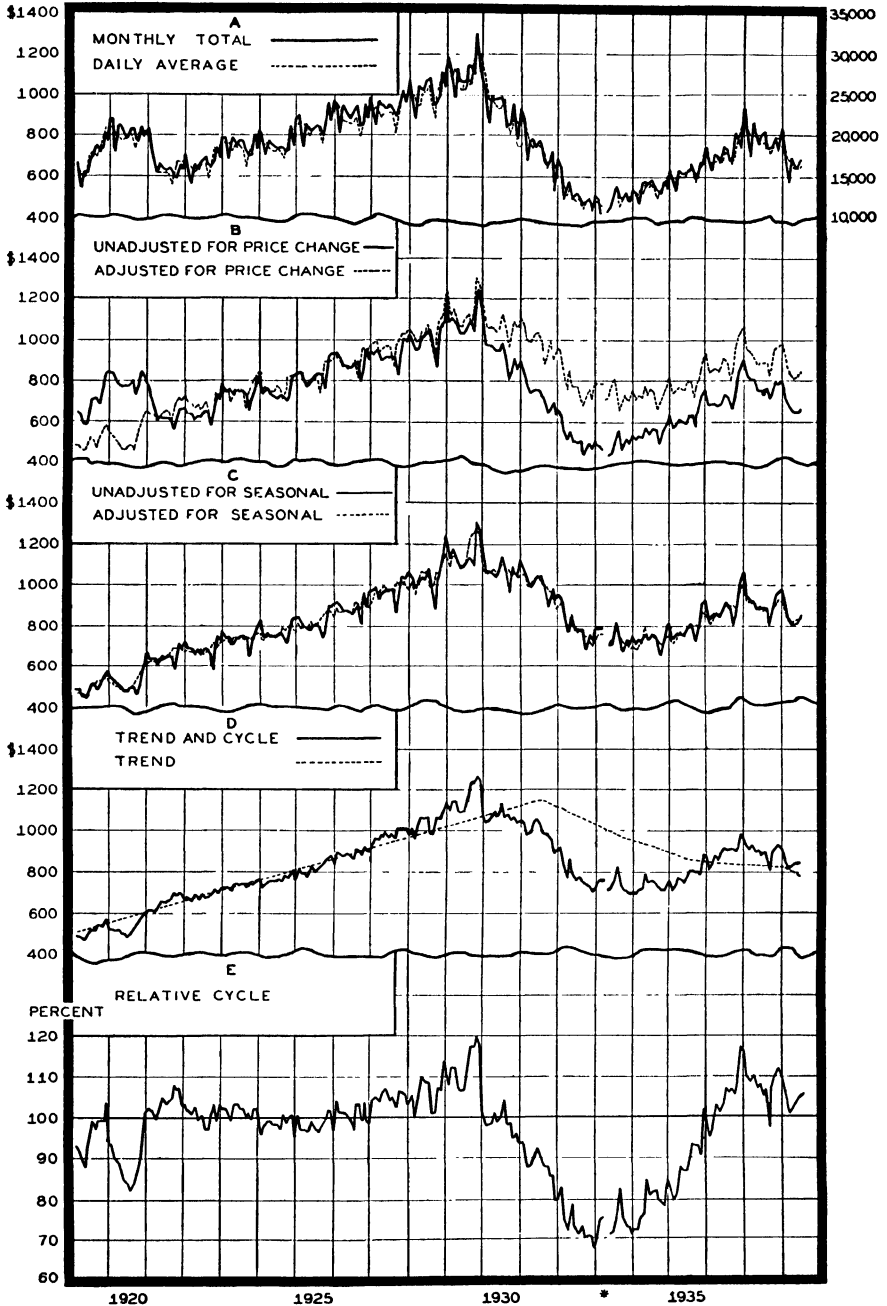
SUCCESSIVE STEPS IN THE ANALYSIS OF A TIME SERIES
BANK DEBITS, MONTHLY OBSERVATIONS 1919-38

DAILY

(000 OMITTED)

MONTHLY

(000 OMITTED)



* Break caused by "Bank Holiday," March, 1933.

the data. When the price index is above 100 the values of the corrected series are less than those of the uncorrected series. When the price index is below 100 the values of the corrected series are above those of the uncorrected series. The result is a series in column 5 which expresses the values bank debits would have attained if prices had remained at the 1926 level throughout the entire period studied.

The solid curve of Figure 96-B represents the daily average bank debits before price adjustment, the broken curve the same after price adjustment. The solid curve is a reproduction of the broken curve from Figure 96-A, column 3 of the table, and the broken curve of Figure 96-B is from column 5. The effect of the adjustment of price changes is most noticeable prior to 1923 and after 1929. Between 1923 and 1929 the price level varied comparatively little from the 1926 base; hence the differences between the uncorrected and corrected curves are less noticeable.

ADJUSTMENT OF SEASONAL VARIATION

The next adjustment to be made depends upon the method selected for measuring seasonal variations. If the ratio-to-trend method were used, the order of removal of the components would be trend and then seasonal. In this analysis the link-relative method is used; hence the order of removal is seasonal and then trend. The link relatives as shown in column 6 of Table 137 are calculated to the end of 1936. The computation of the seasonal index is carried out in Table 138. Column 1 shows the medians of the link relatives, column 2 the medians changed to January as a base, column 3 the fixed base medians corrected for trend residue, and column 4 the adjusted seasonal index. This index is transferred to column 7 of Table 137. The adjustment of seasonal variation is made by dividing each January item of column 5 by the January seasonal index and so on for each month. The resulting series in column 8 is free of seasonal influence and contains trend and cycle. The effect of the correction for seasonal can be seen in Figure 96-C. The solid curve (column 5, Table 137) differs most from the broken curve (column 8, Table 137) in August and December, the months in which debits exhibit the greatest response to seasonal influence. This same smoothing effect is present in other months although less visible on the chart. The effect of the seasonal adjustments is also evident by comparing the solid curves of Parts C and D of the chart.

TABLE 138

COMPUTATION OF SEASONAL INDEX OF BANK DEBITS BY THE LINK-RELATIVE METHOD

MONTHS	(1) MEDIAN OF LINK RELATIVES FROM TABLE 137, COL. 6	(2) MEDIAN CHANGED TO FIXED JANUARY BASE	(3) CORRECTED FIXED-BASE INDEX	(4) ADJUSTED CORRECTED FIXED-BASE INDEX
January	95.60	100.00	100.00	102.81
February	100.05	100.05	99.72	102.53
March	96.40	96.45	95.79	98.49
April	100.10	96.55	95.55	98.24
May	100.30	96.84	95.51	98.20
June	103.15	99.89	98.23	100.99
July	98.45	98.34	96.35	99.06
August	90.50	89.00	86.68	89.12
September	108.55	96.61	93.95	96.59
October	107.20	103.57	100.58	103.41
November	99.45	103.00	99.68	102.48
December	105.60	108.77	105.12	108.08
January	95.60	103.98	100.00	

ADJUSTMENT OF TREND—MOVING-TREND METHOD

The final step in the analysis is the measuring and removal of trend. The method of moving-straight-line trend has been selected for the purpose.²

The moving trend is a measure that possesses the characteristics of a moving average but can be computed for current periods, thus overcoming the greatest disadvantage of the moving average. The computation of a moving trend consists of two steps: (1) the fitting of a straight line to some earlier base period, (2) the projection of this base-period line from period to period, the last value of the projected line being used each time as the current trend value.

The initial step is to determine what length of base period should be used. The longer the period, the more stable the moving trend will be, but a base period of from ten to twenty years is usually satisfactory. There are, however, two practical considerations which usually determine the length of the base period: the first is the number of earlier years for which a given set of data is available, the second is the requirement that a straight line must be a proper description of the trend of the base period.

With this preliminary statement in mind, the discussion can be

² The analysis does not depend upon the use of moving trend. A moving-average, straight-line, or any other type of mathematical curve could be substituted at this point. Columns 9 and 10 of Table 137 would be changed accordingly but the computation of column 11 (relative cycles) would be carried out in the same way.

TABLE 139

COMPUTATION OF TREND OF BANK DEBITS, 1919 TO JUNE, 1938. STRAIGHT LINE FITTED TO ANNUAL DATA 1919-31, MOVING STRAIGHT LINE FITTED TO ANNUAL DATA 1931-37, AND TO MONTHLY DATA JULY, 1937-JUNE, 1938

YEAR	(1) DAILY AVERAGE COMPUTED FROM TABLE 137 COLUMN 5 Y	(2) x	(3) ΔY	(4) $\frac{N-1}{2} (Y_k - N + Y_k)$
1919	502.14	-6	-3,012.84	
1920	520.10	-5	-2,600.50	
1921	651.22	-4	-2,604.88	
1922	683.47	-3	-2,050.41	
1923	740.11	-2	-1,480.22	
1924	767.20	-1	-767.20	
1925	819.91	0		
1926	887.47	+1	+887.47	
1927	977.90	+2	+1,955.80	
1928	1,046.20	+3	+3,138.60	
1929	1,150.89	+4	+4,603.56	
1930	1,059.76	+5	+5,298.80	
1931	982.68	+6	+5,896.08	
$\Sigma Y = 10,789.05$			+9,264.26	
Average	829.93			
kth year or month	Y_k		$\begin{matrix} k-1 \\ \Sigma(vY) \\ k-N \end{matrix}$	
1931.....				
1932	782.33		9,264.26	$6 \times (502.14 + 782.33) = 7,706.82$
1933	735.84		6,684.17	$(520.10 + 735.84) = 7,535.64$
1934	733.89		3,670.67	$(651.22 + 733.89) = 8,310.66$
1935	787.81		1,347.57	$(683.47 + 787.81) = 8,827.68$
1936	897.01		-508.93	$(740.11 + 897.01) = 9,822.72$
1937	903.03		-1,418.09	$(767.20 + 903.03) = 10,021.38$
1937				
July	881.77		-260,592	$77.5 \times (750.81 + 881.77) = 126,525$
August	809.19		-273,767	$(704.84 + 809.19) = 117,337$
September	853.18		-296,307	$(751.31 + 853.18) = 124,348$
October	933.16		-311,894	$(819.05 + 933.16) = 135,796$
November	947.86		-316,067	$(826.83 + 947.86) = 137,538$
December	980.37		-318,605	$(827.21 + 980.37) = 140,087$
1938				
January	870.56		-318,714	$(833.52 + 870.56) = 132,066$
February	833.51		-326,991	$(812.63 + 833.51) = 127,576$
March	807.33		-339,816	$(784.03 + 807.33) = 123,330$
April	811.12		-356,937	$(778.03 + 811.12) = 123,159$
May	820.13		-374,258	$(803.82 + 820.13) = 125,856$
June	842.91		-388,889	$(810.34 + 842.91) = 128,127$

centered on the bank debits series. The best method of determining the length of the base period is from a graph of the data. Study of the solid curve in Figure 96-D indicates that a straight line can be fitted to the period 1919-31, inclusive. The moving trend will therefore be computed on a thirteen-year base.

TABLE 139 (Continued)

COMPUTATION OF TREND OF BANK DEBITS, 1919 TO JUNE, 1938. STRAIGHT LINE FITTED TO ANNUAL DATA 1919-31, MOVING STRAIGHT LINE FITTED TO ANNUAL DATA 1931-37, AND TO MONTHLY DATA JULY, 1937-JUNE, 38

(5)	(6)	(7)	(8)	(9)	(10)	(11)
$\sum_{k-N+1}^{k-1} Y$	$\sum_{k-N+1}^k (xY)$	$\sum_{k-N+1}^k Y$	$\sum_{k-N+1}^k Y$ $\frac{k-N+1}{N}$	$\sum_{k-N+1}^k (xY)$ $\frac{k-N+1}{\sum x^2}$	$\frac{N-1}{2}$ $\frac{k-N+1}{\sum x^2}$	TREND
...	..		.	.	.	524.53
.	...	.	.	.	.	575.43
.	...	.	.	.	.	626.33
..	..	.	.	.	.	677.23
..	..	.	.	.	.	728.13
..	..	.	.	.	.	779.03
....	..	.	.	.	.	829.93
....	..	.	.	.	.	880.83
....	..	.	.	.	.	931.73
....	..	.	.	.	.	982.63
.....	..	.	.	.	.	1,033.53
.....	..	.	.	.	.	1,084.43
.....	..	.	.	.	.	1,135.33
						T_k
...	9,264.26	10,789.05	829.93	+50.90	+305.40	1,135.33
10,286.91	6,684.17	11,069.24	851.48	+36.73	+220.38	1,071.86
10,549.14	3,670.67	11,284.98	868.08	+20.17	+121.02	989.10
10,633.76	1,347.57	11,367.65	874.43	+7.40	+44.40	918.83
10,684.18	— 508.93	11,471.99	882.46	— 2.80	— 16.80	865.66
10,731.88	— 1,418.09	11,628.89	894.53	— 7.79	— 46.74	847.79
10,861.69	— 2,258.40	11,764.72	904.98	— 12.41	— 74.46	830.52
139,700	— 273,767	140,582	901.17	— .87	— 67.42	833.75
139,877	— 296,307	140,687	901.84	— .94	— 72.85	828.99
139,935	— 311,894	140,788	902.49	— .99	— 76.72	825.77
139,969	— 316,067	140,903	903.22	— 1.00	— 77.50	825.72
140,076	— 318,605	141,024	904.00	— 1.01	— 78.28	825.72
140,196	— 318,714	141,177	904.98	— 1.01	— 78.28	826.70
140,343	— 326,991	141,214	905.22	— 1.03	— 79.83	825.39
140,401	— 339,816	141,235	905.35	— 1.07	— 82.93	822.42
140,451	— 356,937	141,258	905.50	— 1.13	— 87.58	817.92
140,480	— 374,258	141,291	905.71	— 1.18	— 91.45	814.26
140,487	— 388,889	141,307	905.82	— 1.23	— 95.33	810.49
140,497	— 401,259	141,340	906.03	— 1.27	— 98.43	807.60

All of the trend computations are given in Table 139. The device of fitting the straight line to annual data and then distributing the annual change to the monthly items has been used to save labor. The initial computations for the years 1919-31 appear in columns 1, 2, and 3. The annual increment is obtained in the usual way,

$$\frac{\Sigma(xY)}{\Sigma x^2} = \frac{\Sigma(xY)}{\frac{N^3 - N}{12}} = \frac{+9264.26}{182} = +50.9$$

The average for the initial thirteen-year period is 829.93, and the annual trend values in column 11 are the result of adding the increment (50.9) successively for years after 1925 and subtracting the increment successively for years before 1925, starting at the average in each case.

The moving trend begins with 1932. The process of fitting a moving trend to bank debits consists in computing successive thirteen-year straight-line trends for the years 1920-32, 1921-33, etc., and using the final values of these "yardsticks" as the trend. For each of these successive thirteen-year periods, the process shown in the upper part of the table could be repeated. However, the labor involved in so many computations of trend may be greatly abbreviated by the use of two formulas. The first formula will provide the value $\Sigma(xY)$ in determining the trend increment for each new position of the thirteen-year yardstick. The second will give the actual moving trend value for the desired year.

$$\text{Formula 1: } \sum_{k-N+1}^k(xY) = \sum_{k-N}^{k-1}(xY) + \frac{N-1}{2}(Y_{k-N} - Y_k) - \sum_{k-N+1}^{k-1}Y$$

$$\text{Formula 2: } T_k = \frac{\sum_{k-N+1}^k Y}{N} + \frac{N-1}{2} \left(\frac{\sum_{k-N+1}^k (xY)}{\Sigma x^2} \right)$$

k = the date of the year to which the trend is to be projected.

$k-1$ = the date of the year from which the trend is to be projected.

N = the number of years in the base period.

Σx^2 = the sum of the squares of the deviations of N years from the middle year as an origin³ and is obtained by the formula, $\frac{N^3 - N}{12}$.

³ The value of Σx^2 remains the same, as long as no change is made in the number of time periods used in computing the moving trend.

$\sum_{k-N+1}^{k-1} Y$ = the sum of the values of the data for $N-1$ consecutive years ending with the year from which the data are being projected.⁴

$\sum_{k-N+1}^k Y$ = the sum of the values of the data for N consecutive years ending with the year to which the data are being projected.

$\sum_{k-N}^{k-1} (xY)$ = the sum of the products of the data times the deviations of the years from the middle year for N years ending with the year *from* which the trend is being projected.

$\sum_{k-N+1}^k (xY)$ = the sum of the products of the data times the deviations of the years from the middle year for N years ending with the year *to* which the trend is being projected.

T_k = the value of the trend for the year to which it is being projected. T_k in this formula, and in Table 139, is the same as Y_t in Table 137.

The application of the formulas is shown in the lower part of Table 139. Formula 2 is used in the first row as a check of the trend value for 1931. The $\sum_{1919}^{1931} (xY)$ (column 6) is $\Sigma(xY)$ from the upper half of column 3. The figure in column 7 is the sum of the daily deflated debits for the years 1919 to 1931 inclusive. This follows from the symbols at the top of column 7 when $k-1=1930$ (the year from which the trend is being projected). Column 8 is the average daily debits obtained by dividing the total in column 7 by 13. Column 9 gives the annual increment of the straight line fitted to the years 1919-31. Column 10 is column 9 multiplied by 6 to give the amount which must be added to the average from column 8 to obtain the value of the trend line in 1931, column 11. The steps carried out in the first row of the lower part of the table are merely an alternative way of obtaining the trend value for 1931 and the result must be identical with that given for 1931 in column 11 of the upper part of the table.

⁴ The subscript below Σ denotes the earliest year included in the summation and the superscript above Σ denotes the most recent year included in the summation.

The use of the formulas to obtain the value for 1932 of the trend fitted to the years 1920–32 is shown in the second row of the lower part of the table. The symbols in the formula have the following meanings,

$$k - 1 = 1931$$

$$k = 1932$$

$$N = 13$$

$$\Sigma x^2 = 182$$

$$Y_{k-N} = Y_{1919}$$

$$Y_{k-N+1} = Y_{1920}$$

and the formulas become,

$$\sum_{1920}^{1932} (xY) = \sum_{1919}^{1931} (xY) + 6(Y_{1919} + Y_{1932}) - \sum_{1920}^{1931} Y$$

$$T_{1932} = \frac{\sum_{1920}^{1932} Y}{13} + \left(\frac{6 \sum_{1920}^{1932} (xY)}{182} \right)$$

Substituting the values in the preceding equations gives,

$$\begin{aligned} \sum_{1920}^{1932} (xY) &= 9264.26 + 6(502.14 + 782.33) - 10286.91 \\ &= 9264.26 + 7706.82 - 10286.91 = +6684.17 \end{aligned}$$

and

$$\begin{aligned} T_{1932} &= \frac{11069.24}{13} + 6 \left(\frac{6684.17}{182} \right) \\ &= 851.48 + 6 \times 36.73 \\ &= 851.48 + 220.38 = 1071.86 \end{aligned}$$

Similar computations for successive thirteen-year periods lead to the trend values in column 11 for the years 1933–37.

Beginning with July, 1937, the method of computing the moving trend on a monthly basis is illustrated. The same formulas are applied, with the same definition of symbols used, except that the unit of time now becomes a month instead of a year, in every case. Thus, instead of thirteen (years) N now becomes 12×13 or 156; k becomes the *month* to which the trend is being projected; $k - 1$ the *month* from which the trend is being projected. For the July, 1937, computation the symbols become,

$$k - 1 = \text{June, 1937}$$

$$N = 13 \times 12 = 156$$

$$Y_{k-N} = Y_{\text{July, 1924}}$$

$$k = \text{July, 1937}$$

$$\Sigma x^2 = \frac{(156)^3 - 156}{12} = 316355$$

$$Y_{k-N+1} = Y_{\text{Aug., 1924}}$$

and the formulas become,

$$\sum_{\text{Aug., 1924}}^{\text{July, 1937}} (xY) = \sum_{\text{July, 1924}}^{\text{June, 1937}} (xY) + 77.5 (Y_{\text{July, 1924}} + Y_{\text{July, 1937}}) - \sum_{\text{Aug., 1924}}^{\text{June, 1937}} Y$$

and

$$T_{\text{July, 1937}} = \frac{\sum_{\text{Aug., 1924}}^{\text{July, 1937}} Y}{156} + 77.5 \left(\frac{\sum_{\text{Aug., 1924}}^{\text{July, 1937}} (xY)}{316355} \right)$$

Substituting from Table 139 gives,

$$\sum_{\text{Aug., 1924}}^{\text{July, 1937}} (xY) = -260592 + 77.5 (750.81 + 881.77) - 139700$$

$$= -260592 + 126525 - 139700 = -273767$$

$\sum_{\text{July, 1924}}^{\text{June, 1937}} (xY)$ was obtained by multiplying the monthly items of

Table 137, column 5, from July, 1924, through June, 1937, by the x deviations, in months, from the middle of the 156 months included and summing the products. The computation which is in the usual form is not reproduced here. The first product is 750.81×-77.5 and the last one is $882.98 \times +77.5$. The values of debits for July, 1924, and July, 1937, come from column 5 of Table 137, and 139700 is the sum of the entries in column 5 from August, 1924, through June, 1937.

The second formula gives the value of the trend for July, 1937,

$$\begin{aligned} T_{\text{July, 1937}} &= \frac{140582}{156} + 77.5 \left(\frac{-273767}{316355} \right) \\ &= 901.17 + (77.5 \times -.87) \\ &= 901.17 - 67.42 = 833.75 \end{aligned}$$

The two formulas are applied in the same way in successive rows of the table to obtain the values of the trend for the several months up to June, 1938.

There are some features of the computation in the lower part of Table 139 which should be called to the reader's attention:

a) Each entry of column 6 is transferred directly to the succeeding row of column 3.

b) Y_k in column 4 is taken directly from column 1.

c) Columns 5 and 7 can be computed simultaneously with the aid of column 4. This can be explained from the figures in the table. Starting with the first entry in column 7

	10,789.05	column 7 row 1
(—)	502.14	column 4 row 2
	10,286.91	column 5 row 2
(+)	782.33	column 4 row 2
	11,069.24	column 7 row 2
	etc.	

The annual trend increment is now transferred from column 11 of Table 139 to column 9 of Table 137. The distribution of the annual trend to the monthly items as shown in column 10 of Table 137 is performed by the method explained in chapter XXIII, page 608, as far as June, 1931. The trend for July, 1931, is obtained by subtracting one-half of the monthly decrement between the middle of 1931 and the middle of 1932 from the 1931 figure. The work follows:

$$(1135.33 - 1071.86) \div 12 = 5.29$$

and

$$1135.33 - (5.29 \div 2) = 1132.69,$$

the value of the trend for July, 1931. The values of the trend for August and succeeding months through June, 1932, are obtained by subtracting the monthly decrease repeatedly. For July, 1932, the process explained for July, 1931, is repeated. The same plan is followed in writing the monthly trend values through June, 1937. Beginning with July, 1937, the monthly trend values are transferred directly from column 11 of Table 139 to column 10 of Table 137.

The calculation of the trend is now complete and its position can be seen in Figure 96-D. It is, of course, a rising straight line from January, 1919, to June, 1931. From that point it moves gradually down in correspondence with the decline of bank debits through 1932, 1933, and 1934. The trend declines more slowly as the curve of the data levels off in 1935 and 1936. The trend remains practically level in 1937 as the curve increases but turns down as debits decline during the first half of 1938. The trend line can be continued as future data become available and its path will correspond in direction to major short-time movements. On the other hand the change in trend will be mild compared to the short-time changes in debits.

THE CYCLICAL FLUCTUATIONS

The trend is removed by dividing each monthly item of column 8, Table 137, by the corresponding trend value in column 10. The results

are the relative cycles expressed as percentages of trend, i.e., the 100 per cent horizontal base line represents the trend.

The cycles are shown in Figure 96-E. As represented by bank debits business activity was about 10 per cent below the average during the first half of 1919 but rose to the average level at the end of the year. A sharp drop early in 1920 was followed by an equally sharp recovery by the end of the year. The curve is above its average through 1921 and to the middle of 1922. From the middle of 1922 to the end of 1926 there is little cyclical fluctuation in the path followed by the curve. The minor variations shown during this period are irregular in character. From the beginning of 1927 to the last quarter of 1929 business activity expanded somewhat irregularly reaching a high point 20 per cent above the average. This expansion is an expression of the boom in general business activity augmented somewhat by the nation-wide participation in security speculation. The decline after October, 1929, was interrupted briefly during the first half of 1930 but subsequently continued to the low point of the depression at the end of 1932. The increase in January and February, 1933, was more financial panic than business expansion and the bank holiday of March followed. After the banks reopened an increase got under way in June and July but was replaced by decline in the later months of the year. After the devaluation of the dollar at the end of January, 1934, bank transactions began a steady expansion which was not terminated until the end of 1936. In 1937, decline was followed by partial recovery and the pattern of the curve was similar during the first half of 1938.

The path followed by the cycles of bank debits describes with some accuracy the movements of general business after 1926 but is faulty for years prior to 1926. There are several possible reasons for the inadequacy of this series during the years 1919-26. (1) Bank debits were first collected in 1919 and it is quite possible that the reporting service was less efficient in the early years. (2) The depression of 1921 apparently was more acute from the point of view of prices than volume of business. (3) The bank debit series is not a sensitive indicator at any time, but this lack of sensitivity appears to have been especially evident during these early years. In particular the series shows only a faint trace of the business expansion of 1923 and the decline which culminated in the middle of 1924.

All in all the curve of Figure 96-E cannot be taken as an adequate representation of business activity since 1919. It is, however, a good

description of the cyclical fluctuations of bank debits. The idea of identifying the cycles of bank debits with those of general business has been brought into this chapter as a forerunner of the section in the next chapter which deals with the reasons why single series such as bank debits may go far afield as indexes of general business conditions.

PROBLEMS

1. *a)* Fit a straight-line trend to the bank debits series of the text.
b) Remove the trend.
c) Compute the index of seasonal variation by the ratio-to-trend method.
d) Compare the results of (*c*) with the seasonal index used in the text.
2. *a)* Complete the computation of the relative cycles from Problem 1.
b) Plot the cycles and compare the results with the cycles in the text.
3. *a)* Compute the index of seasonal variation for bank debits by the moving-average method.
b) Plot on the same graph the seasonal indexes obtained (1) in the text, (2) in Problem 1, (3) in this problem.
c) Discuss the similarities and differences of the three indexes.
4. *a)* Fit a straight-line trend to the yearly averages of Table 139, column 1, from 1919 to 1937.
b) Complete the computation of relative cycles in monthly data.
c) Plot your relative cycles along with those obtained in column 11 of Table 137.
d) Discuss the differences of the two cyclical curves; hence discuss the effect upon the cyclical remainders of the use of different types of trend.

REFERENCES

- BRUMBAUGH, MARTIN A., *Direct Method of Determining the Cyclical Fluctuations of Economic Data*. New York: Prentice-Hall, Inc., 1926.
- HALL, LINCOLN W., "A Moving Secular Trend and Moving Integration," *Journal of the American Statistical Association*, Vol. XX (March, 1925).
- HALL, LINCOLN W., "Determination of Past and Present Secular Trends," *Journal of the American Statistical Association*, Vol. XXI (June, 1926).
- PERSONS, WARREN M., "Indices of General Business Conditions," *The Review of Economic Statistics*. Cambridge: The Harvard University Committee on Economic Research, Vol. I, (January, 1919).

A complete statement of the method of analyzing time series.

CHAPTER XXV

INDEXES OF BUSINESS CONDITIONS

NEED FOR EXTERNAL INFORMATION

THE MANAGEMENT of a business enterprise requires the exercise of judgment of a high order. Success depends largely upon the ability to analyze existing conditions and to evaluate their meaning in terms of the operations of a particular business concern. In increasing numbers business men are attempting to secure numerical facts on which to base decisions. That factual basis consists of two parts: (1) complete internal records of the concern's operations and (2) all pertinent external information available. The most important external information needed is the existing state and direction of change of general business conditions. For many years business men have been demanding more and more information about the movements of general business and of parts of the entire structure. In response to this demand statisticians have developed a variety of measures of business operations. This chapter is devoted to an explanation of the methods employed in preparing such measures.

TWO TYPES OF BUSINESS INDICATORS

For years such series as pig-iron production, bank clearings, and business failures have been used as indicators of the course of business in general. These are known as single-series indicators as distinguished from composite indicators which have been extensively developed since about 1920. A discussion of the characteristics of single-series indicators will provide a background for the major development which deals with composite indexes.

Single Series

Any series drawn from basic production or marketing activity might be taken to reflect business in general, but certain series that are well understood by business men have been widely accepted as indicators. Following are the principal features of some of the commonly used series.

Business Failures.—Since an increase in failures demonstrates financial stringency among business concerns, this series has considerable value in showing the onset of business decline, but is less useful in periods of business recovery since a decline in the number of failures tends to lag somewhat behind other signs of increased business.

Bank Clearings.—Nearly all banks are members of a local clearing house association, through which pass all checks and drafts accepted by each bank to be charged against the individual accounts of depositors in other banks. Since most business transactions are settled by check or draft, the value of these "bank clearings" reflects a large part of daily business activity. The significance of this series is greatly diminished by bank mergers and the growth of branch banks both of which decrease the number of independent banking establishments conducting operations through the clearing house.

Bank Debits.—The defects of the bank clearings series have been partially overcome by the use of a series composed of all debits to individual accounts in banks that are members of clearing house associations. These debits are more representative than clearings because, in addition to transactions between banks, debits include all direct withdrawals from deposits and intrabank clearings in branch banks. Both clearings and debits may expand or contract due to change in prices unaccompanied by change in the physical volume of business.

Bank debits in New York City are used separately because they are weighted heavily with speculative transactions. The series composed of debits in all other reporting cities is ordinarily used as a measure of general business conditions. However, the analysis in the preceding chapter demonstrated rather clearly that at least in the years between 1919 and 1926 bank debits could not be taken to indicate the course of business in general.

Pig-Iron Production.—The smelting of iron ore is the initial stage of the manufacture of iron and steel products so basic in our industrial system. Hence changes in blast-furnace production are presumed to occur in advance of corresponding changes in general business. The practice of stocking pig iron at the beginning of a recession decreases the value of the series as an indicator at the peak of prosperity, and its general value has been diminished somewhat by the increasing use of scrap iron in the manufacture of steel.

Steel-Ingots Production.—The production of steel ingots, the first step of the manufacture of steel products, is usually considered to be

a better general indicator than pig-iron production. Variations in the stock of steel ingots in the yards of producers decrease the value of this series as an indicator. The rate of operations in steel production is so closely allied to the demands of the railroads and the automobile industry that the series may not be entirely representative of changes in the whole industrial structure.

Unfilled Orders of the United States Steel Corporation.—The backlog of orders lost its significance in the late 1920's when hand-to-mouth buying became prevalent in the steel industry; consequently the series was discontinued in 1933. A new series "Shipments of Finished Steel" is now published monthly by the United States Steel Corporation but it is not a sensitive indicator of business changes because the lapse of time between the placing of an order and the shipment of the finished steel may vary from a few weeks to several months.

Price Indexes.—So many price indexes representing different phases of business activity are available that the problem of selecting the proper one for a given purpose is difficult, and the more comprehensive the index, the less it reflects changes in business. Some of the published indexes containing a few commodities whose prices are sensitive to business changes are better indicators of short-run business activity than the more representative indexes constructed from a large number of commodities. As competitive prices have been replaced in recent years by prices which are managed in one form or another, movements of the price structure are often not directly related to business expansion and contraction.

Security Prices.—Many indexes of prices of stocks representing all types of production and marketing are available in separate groups or combined. The trading of securities in organized markets is usually assumed to be at prices which are the estimate of buyer and seller as to what future profits of business will be. In so far as such estimates are accurate the stock market is a satisfactory indicator of changes in business. There are, however, many cross-currents represented in the prices of securities: investment is fused with speculation, foreign buying or selling is frequently unrelated to business conditions in the United States, the automatic checks and balances of trading are partially superseded by rulings of the Securities and Exchange Commission, and the increased participation of the small operator in the market has reduced the precision with which the market performs its function of anticipating future business activity.

Railroad Freight Car Loadings.—The number of loaded cars moved by the railroads is an indicator of the volume of goods of various kinds going into industry and trade. Although this is one of the best single series indicators of increases or decreases in business activity, there are certain defects to be noted. A loaded car may vary from a few tons to one hundred tons; hence there may be a change in volume of business without any change in number of car loads transported due to the natural tendency to load cars heavier when business expands and lighter when business declines. The increasing percentage of total freight that is being carried by other transportation agencies has also affected this series.

Electric Power Production.—This series is more closely allied to manufacturing than to business in general. Any expansion of manufacturing activity will be reflected in increased consumption of current. The series is gradually becoming more representative as a greater number of manufacturers install electrically driven machinery. But the shift to electricity from other fuels gives this series a strong upward trend; consequently it tends to underemphasize business decline and overemphasize business expansion.

Building Contracts.—New construction is undertaken in response to actual or anticipated expansion of business activity. Declining business finds quick expression in sharp contraction of building operations. Since 1932 public construction has had too much influence on the total value of contracts awarded, thereby reducing its relation to business conditions. The value of contracts awarded is a mixed indicator because it represents partly consumption goods (residences), partly producers' goods (commercial and industrial buildings), and partly public works.

Summary.—If proper allowance is made for special circumstances any one of these series can be used successfully as a guide to the course of business activity. But the great danger lies in not being able to make the proper allowance at any given time; consequently business men usually follow several series simultaneously in an attempt to arrive at an evaluation of the business situation. But this is a cumbersome process at best, requiring intimate knowledge of the various circumstances affecting each series, and withal does not resolve the danger that the significance of changes in individual series will be misjudged. The alternative is to use a combination of a number of individual series, that is, a composite index.

Composite Indexes

When the total business situation is represented by the combination of a number of individual series, the result is known as a composite index. The great advantage of the composite index lies in the fact that the series included can be so selected that they will represent specific elements of the business structure. That is, in a composite index steel-ingot production would be included to measure conditions in the steel industry only, whereas if steel-ingot production is used as a single indicator there is always the tacit assumption, "business follows the steel industry."

The first step in planning a composite index is to determine the purpose for which it is to be used. This purpose will be the ruling consideration in the selection of the series included and the importance assigned to each.

CONSTRUCTION OF COMPOSITE INDEXES

The preparation of a composite index is an application of the methods of index number construction developed in an earlier chapter. The problem discussed there was how to combine a number of individual prices, quantities, or values to obtain an average figure. The parallel problem here is how to combine a number of series pertaining to individual lines of business to obtain an index describing business as a whole. There is, however, one major difference which explains why the construction of business indexes was deferred until the analysis of time series had been explained. A business index usually measures the cyclical rhythm only; hence the other components of each series must be eliminated.

Removal of Other Components

The objection to combining series of original data without any refinement should be fairly obvious from the chapters dealing with the analysis of time series. Different series will exhibit a variety of seasonal movements which may or may not cancel each other when the series are combined. Some series have positive trends, others negative. Even though all of the trends change in the same direction their rates of change will vary, and either situation interferes with the measurement of cycles. Those series which are expressed in dollars will include the effect of price change; in others expressed in physical units this

component will be absent. The proper procedure is to follow the usual steps of correcting for calendar variation, price change where necessary, seasonal variation where present, and trend. The cyclical fluctuations which result from the analysis are in similar form for all series and contain only a component which arises from a common cause and manifests itself in the same way for all series.

Combining Individual Series

The composite index is made by combining the cyclical components of a number of individual series. The selection of the series to include in an index rests primarily on the question of availability, but among available series several factors require detailed consideration as criteria.

Lag.—The cyclical movements may not occur simultaneously in all series. Some respond more quickly than others to an expansion or contraction of business. Stock prices for example are very sensitive to changes in business conditions, while interest rates on commercial paper remain low through the early stages of a business recovery and usually reach their highest level after business has begun to decline from the peak of prosperity. Variations of this kind in the occurrence of the cyclical swings of different series are well known to business men. In fact it is quite unusual for the cyclical movements of a number of series to coincide in their time of occurrence. So long as the series alternate in the time of their cyclical movements there is no objection to combining them to obtain a composite cycle. For example, three series A, B, and C may reach a peak of prosperity in the order B, C, A; they may reach the bottom in the following depression in the order C, A, B, and return to prosperity in the order A, B, C.

If, however, the cyclical movements of one or more series consistently lead others or lag behind others, such series should not be combined with the others in a composite index without making some adjustment for this lead or lag. Unless this correction is made the nonconcurrent cycles of the individual series will conceal the cyclical movements of the composite index.

Methods of Testing for Lag.—Series intended for use in a composite index should be tested for lag after the cyclical component has been computed. There are three methods available for determining the existence of lag: (1) tabulation of turning points of the individual cyclical curves, (2) comparison of graphs of the individual cyclical

curves over a light table,¹ (3) computation of coefficients of correlation between the individual cyclical fluctuations in pairs.

Tabular: This method consists of tabulating in parallel columns the dates of the turning-points of the cycles of the several curves. If the dates for any curve are consistently earlier than the others, it is said to lead them. If the dates are consistently later than the others, it is said to lag behind them. Table 140 gives a tabulation of the turning-points of the cycles of an index of industrial stock prices and interest rates on four-to-six month commercial paper from 1919 to 1937. These turning-points have been read from Figure 97.² The

FIGURE 97

RELATIVE CYCLES OF INDUSTRIAL STOCK PRICES AND COMMERCIAL PAPER RATES, MONTHLY DATA, 1919-37

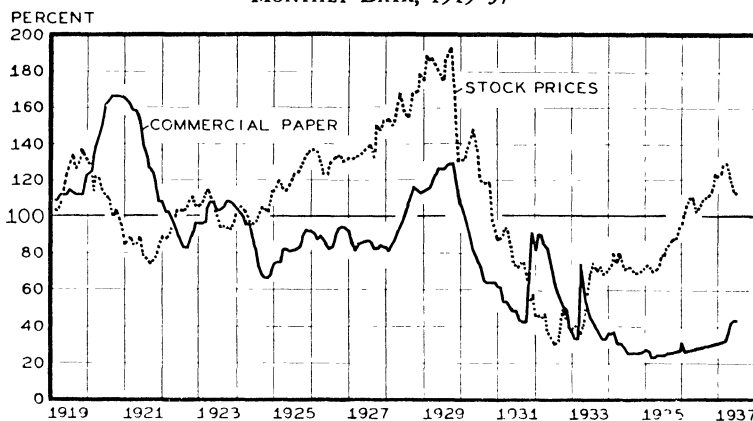


TABLE 140

LIST OF TURNING-POINTS OF CYCLES OF AN INDEX OF INDUSTRIAL STOCK PRICES AND INTEREST RATES ON FOUR-TO-SIX MONTH COMMERCIAL PAPER, 1919-37

DATE OF CYCLE FROM PEAK TO PEAK	TURNING- POINT	INDEX OF INDUSTRIAL STOCK PRICES	INTEREST RATES ON COMMERCIAL PAPER
1919-23.....	{ Peak Trough Peak	Oct. 1919 Aug. 1921 Mar. 1923	Sept. 1920 Aug. 1922 Sept. 1923
1923-29.....	{ Trough Peak	Oct. 1923 Sept. 1929	Oct. 1924 Oct. 1929
1929-37.....	{ Trough Peak	June 1932 Mar. 1937	Mar. 1935

¹ A light table is usually as high as an ordinary desk and the dimensions of the top are about 1½ x 2 feet. The one-piece glass top has a light or lights directly under it, so that charts placed on the glass become transparent. Thus two or more curves on separate sheets can be compared.

² The computation of the cycles of the two series is not reproduced because of lack of space.

cycles of commercial-paper rates lag about one year behind those of an index of stock prices according to the evidence of the years prior to 1929. After 1929 the usual relation between the cyclical movements of the two series does not hold, but the lag of interest rates is still apparent.

If additional series were included in the comparison, they would be tabulated in Table 140 in the same way. The comparisons could then be made in pairs or for several series simultaneously. The dates recorded in the table could be read from the computation sheets without the aid of a graph, if necessary, but the cyclical movements are usually more easily determined from a graph.

Graphic: The second method of determining the existence of lag consists of constructing a separate graph of the cyclical movements of each series which is to be studied. These graphs are placed on a light table in pairs and moved forward and backward until the cycles of the two correspond. The comparison of the graphs discloses the presence of any series whose movements consistently lead the others or lag behind them and the number of time periods of lead or lag involved.

The graphic method is, of course, approximate, because it requires that a judgment be made merely by inspection of the relation between two sets of cycles covering a period of years. In some cases the relation is fairly obvious but, when only one or two months of cyclical difference is involved or the cyclical movements are quite irregular, considerable uncertainty may arise as to the existence of lag. Usually it is desirable to have more than one person study the graphs independently and then compare the conclusions of the different observers.

Correlation: Although the existence of lag can usually be detected by either of the methods previously explained, the exact period of lag is less easily determined. Frequently it is desirable to compute coefficients of correlation between pairs of cyclical fluctuations using several different periods of lag, the proper period being indicated by the maximum correlation. The details of this process are explained in chapter XXVII, pages 742-44.

The best procedure in determining the existence and amount of lag is to use tabular or graphic analysis as a preliminary step to get an approximate result and then if feasible apply correlation to make the final determination. In some cases, however, the cyclical movements of two series are so erratic that the correlation between them never becomes large enough to be interpreted even though an irregular lag

is present. In such cases the use of correlation must be abandoned and reliance placed on tabular and graphic methods. The extent to which lag should be studied depends in a particular case upon the degree of refinement used in constructing an index, the existence of erratic cyclical fluctuations, and the length of the period for which data are available. The latter point is important, because lag established from data for a short period may prove to be inaccurate.

Finally the maker of an index should be constantly on guard for changes in the economic relation between series that might result in a change in the lag between them. This type of change is apparent in Figure 97 since 1933. The cyclical movements of interest rates have not corresponded to those of stock prices due to the many changes in bank operations since that date.

Inverted Series.—There are certain phenomena in our economic system which manifest themselves by inverse cyclical movements, that is, they go down when business goes up and go up when business goes down. Examples are failures of business concerns, the ratio of bank reserves to deposits, unemployment compensation, vacancies in urban dwellings, and public construction. The inverse movements of these series have definite causes. A decline in business volume causes marginal concerns to fail in increased number, while expanding business volume with rising profits permits the survival of weak business concerns and decline in failures. The decline in bank loans following a business crisis is more rapid than the contraction of the currency, resulting in an increased ratio of reserves to deposits in the banking system, but expanding loans precede currency expansion in business revival thereby increasing deposits and reducing the ratio of reserves to deposits. Declining volume of production and distribution of goods causes contraction of employment thereby swelling unemployment rolls and increasing the demand for unemployment compensation, and expanding business produces the reverse result. Depression causes families to double up or to dissolve, decreases the marriage rate, and tends to drive families out of urban centers thereby increasing the number of vacant dwellings in cities despite an attendant decline in residential construction. Business expansion sets opposite forces in motion augmenting the demand for dwellings in advance of increased residential construction and causing a reduction in the number of vacant dwellings. Long-term planning of permanent public improvements is based upon the assumption that public spending should be increased during depres-

sion to take up the slack in private industry, and conversely that public spending should be curtailed during prosperity when private industry needs no assistance.

The inclusion in a composite index of inverted series such as those described above requires an inversion of their cyclical movements, a simple arithmetic process. If the relative cycles are expressed as per cents above and below zero, the inversion consists simply in changing the signs; if the relative cycles are expressed above and below 100, all relatives above 100 are written an equal amount below 100 and all relatives below 100 are written an equal amount above 100. For example, if unemployment rose to 133 per cent of the average during a depression, this positive variation would be changed to a negative form by expressing it as 67 per cent. After this adjustment the inverted series may be used in the composite index the same as any other.

Variation in Amplitude.—As explained in chapter XXI, page 544, some series have cycles of high amplitude, others low. If several series of each kind are included in a composite index those with cycles of high amplitude will affect the average more than those with cycles of low amplitude. But low-amplitude series often represent more fundamental movements of the business structure than those with high amplitude; hence it is necessary to introduce some device for equalizing cyclical movements of different amplitude before combining the several series.

Method of equalizing amplitude: High or low amplitude is of course a question of the amount of dispersion present in the cyclical fluctuations of a series. Hence amplitude can be equalized by the use of a measure of dispersion. The general procedure is: (a) compute the average deviation or standard deviation for each series for whatever period the data are available, (b) express the cyclical fluctuations in units of the dispersion or assign to each series a weight which is the reciprocal of its dispersion. The importance of cycles of high amplitude is thereby reduced and the importance of cycles of low amplitude is increased, thus producing the desired equalization.

This method does not equalize the amplitudes of the cycles of the several series for each unit of time (i.e., month or week) but equalizes them for the average relation existing during the entire interval for which the dispersion was computed. At any particular time the importance of the several series may vary considerably and on that account the process of equalizing amplitudes has sometimes been criticized.

These criticisms disregard the fact that so long as the original relation of the several dispersions continues to hold no one series can maintain excessive importance for a long period. Without the equalizing step series of high amplitude are certain to maintain excessive importance and series of low amplitude are equally certain to maintain a smaller than proportional importance.

Weights.—After series have been equalized, weights must be applied to them in order to give each the effect it should have due to its relative importance in the business structure. For example, if bank debits and mortgage loans of savings and loan associations are to be combined in the same index, the cycles of the two must first be equalized because the mortgage loan cycles will fluctuate much more than those of bank debits. On the other hand, the cyclical movements of bank debits are of greater importance in measuring business in general than are those of mortgage loans; hence debits must be given greater weight in the index.

The weight to be assigned to each series will depend upon the purpose of the index and the number of series included. In establishing weights there are no rules which can be followed except the very broad one that the weights should be set by someone who has adequate knowledge of the interrelations of the business activities represented by the several series and who is thoroughly familiar with the methods of analysis employed in the construction of the index. This step is usually called the introduction of judgment weights.

The two steps, equalizing amplitudes and introducing judgment weights, can be performed separately but more commonly they are combined in a final set of factors which reflect the net effect of both influences. The combined weighting procedure is illustrated in the Annalist Index of Business Activity described later in the chapter.

Form of the Final Index.—The final index is obtained by combining the individual series properly weighted. Two distinct forms of final index are found in print: (*a*) cyclical indexes, (*b*) trend-cycle indexes. The names describe the components present in the two types of indexes. Trend-cycle indexes in turn are of two kinds: (1) those with trends fitted to individual series, (2) those with trend fitted to the composite index. Each type of index will be explained separately.

The cyclical index: The cyclical values of each series should be expressed as relatives of trend. The average of these weighted relatives for each time period is the combined cyclical index. The individual

relatives and correspondingly the combined index may be expressed as per cent relation to trend or per cent difference from trend. The former leads to an index with values above and below 100 as a base, while the latter leads to an index with values above and below 0 as a base. In either case the horizontal base is a representation of the composite trend of the series included in the index. Positive amplitudes of the composite series portray prosperity and negative amplitudes portray depression.

The distinguishing feature of this method is the representation of the composite trend of the several series as a horizontal base line for the cyclical fluctuations. In this form no idea is conveyed of the path actually followed by the trend.

The trend-cycle index: In recent years it has become desirable to show cycles above and below the actual path followed by the trend instead of in relation to a horizontal line. So long as trend could be properly measured by an increasing straight line, and particularly if the trends of the series included in an index increased at about the same rate, the representation of the composite trend as a horizontal line caused little difficulty to either statistically trained or lay readers. Since 1930 these conditions have ceased to exist; the trends of the postwar decade no longer apply in rate of growth and in some cases not even in direction. All sorts of devices have been employed to secure satisfactory trends. Some of these have been more successful than others, but regardless of the efficacy of the devices their variety has created a situation in which the interpretation of any index of business conditions depends to a large extent upon the method employed in measuring trend. This being the case, the maker of an index is faced with the necessity of showing the actual path followed by the composite trend of his index rather than allowing the trend to stand as a horizontal base for cyclical fluctuations. Trend-cycle indexes may be constructed either by reinstating trend in a cyclical index or by an entirely different process in which trend is not removed from the individual series.

Trend Reinstated: The trends of the several series of a business index will usually be expressed in different kinds of units such as dollars of bank debits, barrels of flour, persons employed, etc. Before such items can be combined they must be changed to per cents. Hence the first step is to select a single year or several years as a base. Each trend is then expressed as a per cent of this base year or years. The com-

posite trend is a weighted average of these per cents. The set of weights should be those used in obtaining the composite cyclical index, or a variation of them which places less emphasis on the amplitude factor and more on the importance factor. The trend-cycle index is obtained by multiplying³ each value of the trend index by the corresponding value of the cyclical index, the latter being expressed in the form of per cent relation (base = 100). The final index numbers are thus expressed as relatives above and below the composite trend.

Trend Not Removed: This method proceeds with the adjustment of calendar variation, price change, and seasonal fluctuation just as in the preceding method, but differs in that the trends are not removed from the individual series. The composite index, constructed in the form of a relative of weighted aggregates, contains trend and cycle and is based on a single year or several years as 100 per cent. A trend may or may not be fitted to the final index. The Babsonchart, described later, is an example of this type of construction.

Three differences should be noted between an index in which trend is reinstated and one in which it is not removed. The cyclical movements of the individual series can be studied in the former but not in the latter. Weights may be adjusted to the amplitude of cyclical fluctuations and the importance of a series in the former whereas they can be related only to importance in the latter. The measurement of composite trend is explicit in the former; such measurement is merely implicit in the latter.

EXAMPLES OF THE CONSTRUCTION OF COMPOSITE INDEXES

Anyone wishing to study the current position of business or the levels of business over a period of years has at his command a number of published indexes, such as the Annalist Index of Business Activity, the Babsonchart of Business Conditions, the Harvard Index of General Economic Conditions, the Index of Industrial Production of the Board of Governors of the Federal Reserve System, and the New York Times Weekly Index of Business Activity.⁴ The number of such indexes gives some indication of the demand by the business man for this type of statistical work.

³ So long as the values of the composite trend remain in the vicinity of 100, the relative cycles may be added instead of being multiplied without substantial loss of accuracy in the results.

⁴ A more complete list of Business Indexes and original sources is included in the references at the end of the chapter.

Various methods of construction are used in the different indexes and each one is intended to measure certain features of business activity. A complete understanding of the several indexes would require a very thorough study. The most that can be done here is to explain a few of the indexes as a guide to the analysis to which any index should be subjected by a person expecting to make practical use of it. The Annalist Index and the Babsonchart have been selected for explanation because they represent two extremes in method of analysis and construction.

The Annalist Index ⁵

The general plan of construction of the Annalist Index of Business Activity is to analyze separately each series included and form the final index as a weighted average of the several individual cyclical components. The result shown in Figure 98 is intended to be a representation of the course followed by the business cycle. The horizontal base line serves as an average level above and below which positive and negative cyclical fluctuations of business are portrayed.

This index has been published continuously since November, 1925, and has been revised twice (1933 and 1936) in the interim. The changes introduced in the 1936 revision have been used in recomputing the index back to 1923; hence no reference to earlier forms of the index is necessary to obtain a continuous record since that date.⁶

Series Included.—The fifteen series used in preparing the combined index are listed in Table 141. Each of these series introduces a different phase of business activity into the index. Production of pig iron and steel ingots represents heavy machinery, tools, and other iron and steel products. Production of lead and zinc represents the mining industry. Mill consumption of cotton, wool, silk, and rayon represents the textile industries. Production of lumber and cement represents the building industry. Production of automobiles, and production of boots and shoes represent, respectively, semi-durable and non-durable

⁵ Publication of *The Annalist* as a separate magazine was discontinued in November, 1940, but several of its features, including the Index of Business Activity, were taken over by *Business Week*. Current data for the Index appear in *Business Week*, but data for months prior to November, 1940, must be obtained from files of *The Annalist*.

⁶ A complete description of the present form of the index appears in *The Annalist*, Vol. 47, No. 1223 (June 26, 1936), pp. 939–43. Previous descriptions of the index appeared in *The Annalist*, Vol. 42, No. 1074 (August 18, 1933), pp. 213 and 238, and Vol. 29, No. 733 (January 28, 1927). These references are the source of the information presented here.

consumers' goods. Electric power production and two series of freight car loadings represent whatever activities of production and distribution of goods, respectively, are not specifically brought into the composite index by the other twelve series. It is evident from this statement that the purpose of the index is to show the major movements of business activity.

Removal of Components Other than Cyclical.—*Calendar variation:* The monthly production or consumption as reported for each series is reduced to an average daily basis to adjust for variations in number of days in the month. The compilers studied each series separately to discover the number of days worked each month since 1920 in the industry represented by that series. The number of days finally selected for each series was necessarily a compromise because of differences in the length of the working week in different parts of the country as well as variations in the holidays celebrated.

Price factor: All of the series in the index are expressed in physical measurement units; hence no correction for price change is required.

Seasonal: The link-relative method is used to establish a moving seasonal pattern for each series. In all of the series except automobile production the seasonal pattern used in any year is based on the nine preceding years; thus the patterns for 1939 are determined from the years 1930–38, inclusive. Prior to 1934, the pattern for automobile production was determined from the seven preceding years. Since 1934, the time of the change in date of introducing new models, the pattern has been established arbitrarily. The seasonal influence is removed from each series by dividing the daily average figure for each month by the seasonal index for that month.

Trend: The problem of measuring the trend of the several series is the most difficult step in the construction of the index. The difficulty arises from the inflexibility of trends measured by mathematical equations. This question was discussed in considerable detail in the chapter dealing with trend. The solution proposed by the compilers of the Annalist Index leads in some series to a result which resembles that obtained by the use of a moving trend, but the actual computation employed is entirely different. The series that required the greatest consideration were pig-iron production, electric power production, freight car loadings (two series), and cotton consumption. Preliminary study indicated that the trend of each of these series had undergone changes since 1929 that made previous methods of measurement

unsuitable and required the establishment of some standard for determining the actual path followed by the trend in recent years. Search for a standard led to the selection of steel-ingot production. It will be necessary, therefore, to explain how the trend of steel-ingot production was determined before proceeding with the five troublesome series.

A straight line was fitted to steel-ingot production for the period 1919-31, and this straight line was projected into subsequent years. The actual production figures were then expressed as per cents of this trend line, and the validity of these cycles was tested by comparing them with the cycles of the series "Ratio of Steel Ingot Production to Capacity of Steel Mills." This series is assumed to contain no trend; hence the cycles can be obtained by determining a normal ratio. This normal was found to be 69 per cent of capacity. Accordingly a set of relative cycles was worked out by dividing the actual ratio by 69 per cent. The two sets of cycles of steel-ingot production (one after removal of trend, the other containing no trend) proved to be "virtually identical." Therefore a straight line fitted to steel-ingot production for 1919-31 and projected into subsequent years appeared to be an accurate description of the growth of the series, and the cycles computed from this trend "were selected as a standard by which to gauge cyclical deviations from normal in other business indicators." ⁷

The determination of the trend of pig-iron production "has been accomplished by dividing average daily pig iron production, seasonally adjusted, by the adjusted index of steel ingot production. This process provides a series of figures which confirms the accuracy of the former long-term trend of pig iron production (a straight line fitted to the period 1921-31) over the period 1919-27 and indicates that a leveling-off process began in 1927. The revised adjusted index of pig iron production is, therefore, based on the old trend line up to the latter part of 1927, whence it continues as a horizontal line at a level of 90,000 tons per day." ⁸

The method of calculating the trend of the two car loadings series (miscellaneous and all other) "is similar to that used in calculating the pig iron index, except that an allowance has had to be made for the different amplitudes, or characteristic widths of fluctuations, in

⁷ *The Annalist*, Vol. 47, No. 1223 (June 26, 1936), p. 939.

⁸ *Ibid.*, p. 939.

steel ingot production and car loadings. Specifically, normal amplitudes were determined by calculating average annual deviations of the seasonally adjusted averages from the mean for each year. The approximate positions of the trend lines were then located by dividing the average daily loadings figures, seasonally adjusted, by the index of steel ingot production after reducing the amplitude of the steel index by the amount indicated in each instance by these differences in amplitude.”⁹

Comparison of the cycles of electric power production with the cycles of steel-ingot production leads to a trend for electric power production by the following procedure. “By allowing for the fact that steel ingot production fluctuates about five times as widely as electric power production, we have determined a new normal for electric power production by dividing the seasonally adjusted power daily averages by the steel index. The trend line thus computed, after being smoothed graphically” shows that “the tendency now is for electric power production to increase at an annual rate of about 6 per cent, as compared with the pre-depression rate of increase of about 10 per cent.”¹⁰

The cycles of cotton consumption do not correspond in either amplitude or period with those of steel-ingot production; consequently a different approach had to be found in determining the trend of cotton consumption. “The percentage of capacity operated by the cotton textile industry affords a basis for estimating normal activity in terms of average daily mill consumption of raw cotton. These figures are available back to 1922, since when the average rate of operation has been 92.3 per cent. We accept that as normal in terms of per cent of capacity. We translate it into terms of average daily cotton consumption by dividing the average daily consumption figure for each year by the ratio of per cent of capacity operated to normal. This yields a consistent series of annual figures which when smoothed, afford the long-time trend line.”¹¹

Various devices were employed in establishing valid trends for the other series of the index. These devices, however, are less complicated and have less general importance than the methods described in preceding paragraphs. Consequently we will be content here with a list of the trends used in the other series.

⁹ *Ibid.*, pp. 939–40.

¹⁰ *Ibid.*, Vol. 47, No. 1220 (June 5, 1936), pp. 830, 831.

¹¹ *Ibid.*, Vol. 47, No. 1223 (June 26, 1936), p. 940.

INCREASING TREND

- Rayon consumption.....projection of straight line fitted to years 1923-35.
- Cement production.....projection of straight line fitted to years 1919-32.
- Zinc production.....projection of straight line fitted to years 1899-1910.

NO TREND¹²

- Wool consumption.....average consumption during years 1919-31 gives value of horizontal line.
- Boot and shoe production...average production during years 1919-31 gives value of horizontal line.
- Lead production.....average production during years 1930-31 gives value of horizontal line.
- Lumber production.....average production during years 1929-31 gives value of horizontal line. A horizontal projection of the average production during 1919-21 was used from 1919 through 1924. The end of 1924 was joined to the beginning of 1930 by a line with negative slope.

COMBINATION INCREASING TREND AND HORIZONTAL LINE

- Silk consumption.....straight line fitted to years 1920-32 used to end of 1930, from whence it is projected as a horizontal line.
- Automobile production.....straight line fitted to years 1919-27 used to middle of 1926, from whence it is projected as a horizontal line.

The trend is removed from each series by dividing the seasonally corrected daily average figure for each month by the corresponding trend figure. The results are relative figures expressed in the form of index numbers. They contain the cyclical component and whatever irregularities remain after the removal of the other components.

Combining the Series.—Each of these series represents a fundamental movement of business activity; therefore there is no lag in

¹² Study of graphs of these series led to the conclusion that they contained neither increasing nor decreasing trends; hence the average value of the several series of data for the periods indicated was used as a base or horizontal "trend" for measuring cycles.

the cyclical fluctuations of any of them nor are any of the cyclical fluctuations inverted.

Varying amplitude: Some of these series have cycles of high amplitude, others low. The method used in measuring amplitudes is described in *The Annalist* as the "average annual percentage deviation of high and low adjusted indices from mean daily averages." Specifically two steps are involved in measuring amplitudes: (1) for each series the difference between the high and low cyclical relatives was computed for each year from 1921 to 1935 (shorter periods had to be used for some series); (2) the average of these annual range figures was computed for each series. These averages appear in column (b) of Table 141 and their use to equalize amplitudes is explained in the next paragraph.

TABLE 141

WEIGHTING SYSTEM USED IN COMBINING THE SERIES OF THE ANNALIST INDEX OF BUSINESS ACTIVITY*

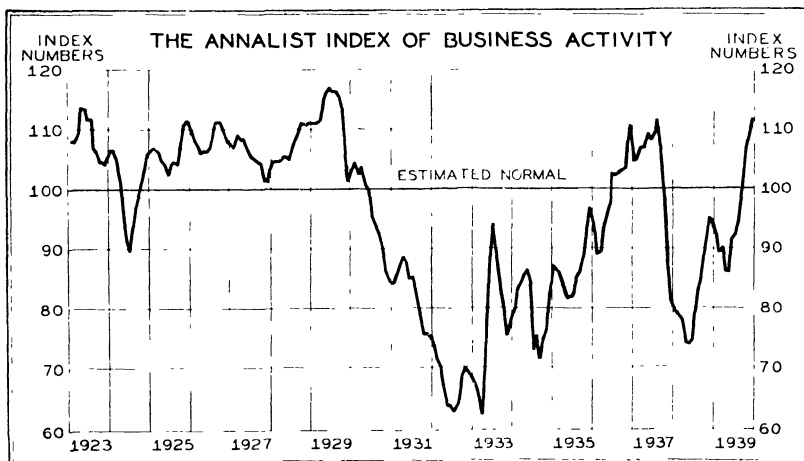
	(a) EFFECTIVE WEIGHT	(b) AVERAGE RANGE	(c) (a) ÷ (b)	(d) ADJUSTED WEIGHTS
Miscellaneous loadings	12	14	.86	.16
Other car loadings	6	13	.46	.08
Electric power production	12	6	2.00	.37
Steel ingot production	20	38	.53	.10
Pig-iron production	10	35	.29	.05
Cotton consumption	10	31	.32	.06
Wool consumption	3	44	.07	.01
Silk consumption	1	35	.03	.01
Rayon consumption	2	51	.04	.01
Boot and shoe production	4	26	.15	.03
Automobile production	8	56	.14	.03
Lumber production	3	23	.13	.02
Cement production	2	24	.08	.01
Zinc production	4	21	.19	.04
Lead production	3	30	.10	.02
Total	100		5.39	1.00

* *The Annalist*, Vol. 47, No. 1223 (June 26, 1936), Table VI, p. 940.

Weights: The complete weighting system is reproduced in Table 141. "The adjusted indices for each series are combined into a composite index by computing weighted averages. Each component is weighted (column a) according to its relative importance and reliability as a business indicator. Iron and steel production and freight car loadings are assigned the heaviest weights because of their universally recognized importance and reliability. Electric power production is also given a heavy weight, but it is weighted slightly less than car loadings because it is somewhat susceptible to temperature and weather

changes. Textile fiber consumption and boot and shoe production are given a combined weight of 20 per cent, thus giving adequate representation to the leading industries concerned with consumers' non-durable goods. Automobile production, with a weight of 8 per cent, is perhaps not given the influence on the composite which its importance at the moment seems to justify; but the value of the motor production series as a general business indicator is reduced by its recurring erratic behavior and the impossibility of computing accurate seasonal indices. Lumber and cement production, and zinc and lead production, are included in order to give effect to the influence of cyclical fluctuations in new construction and mining activity respectively."¹⁸ Column (b) contains the amplitude measures described in the preceding paragraph. Variations in amplitude are adjusted by dividing each weight of column (a) by the amplitude figure of column (b) to give the combined weights of column (c). The weight of a series such as steel-ingot production having cycles of high amplitude is reduced by this division, while the reverse occurs with a series such as electric power production having low amplitude cycles. Column (d), giving the final weights, was obtained by dividing each item of column (c) by the total of the column (5.39).

FIGURE 98
ANNALIST INDEX OF BUSINESS ACTIVITY



Reproduced from *The Annalist*, Jan. 25, 1940 (earlier years added), with the permission of the *New York Times*.

¹⁸ *Ibid.*, p. 940.

Form of the final index: The index for each month is obtained by computing an average of the weighted values of the relative cycles of the fifteen series. The published form of the index is reproduced in Figure 98. The reference line (100) is the composite trend of the fifteen series represented as a horizontal base for positive and negative cyclical fluctuations. The final index is a representation of the cyclical fluctuations of business in the United States. *The Annalist* does not compute a trend-cycle index.

The Babsonchart

The Babsonchart differs from the Annalist Index in several important respects. The Babson Index is computed as a relative of weighted aggregates, the Annalist Index as a weighted average of relatives. The Babson Index is expressed as percentages of a fixed base period, the Annalist Index as percentages of trend as a base. Trends are not removed from the several series in the Babson Index whereas they are removed in the Annalist Index. The final form of the Babsonchart includes a composite trend, but the final form of the Annalist Index has no representation of trend. The Babson Index contains a large number of series and depends partly on inclusiveness to make the results representative, while the Annalist relies on a smaller number of selected indicators to secure representativeness.

One of the earliest in the field, the Babson Business Index has been published continuously since 1905. At various times since then changes in the series included and in the method of construction have been necessary to keep abreast of new developments in the business structure and in the field of statistics, but throughout the entire period the Babsonchart has retained its essential characteristics and its basic principle of cyclical representation, "action is equal to reaction." The following description deals solely with the present form of the index which is constructed to conform to six requirements.

1. It must include as many major phases of business activity and industrial divisions as can be adequately and reliably represented by the series available.
2. It must be adjusted for seasonal variation so that any month may be compared directly with any other month.
3. It must give each series its proper importance in the index. This importance must be determined on the basis of the "value added" by the industry represented by the series.

4. It must adjust for the shifting relative importance of constituent items and of the major groups during the period covered by the chart, and it must provide adequately for incorporating such changes in the future.
5. It must show the actual index of activity—not merely the deviations from some unrevealed “normal,” so that the long time trend of activity can be seen and comparisons made between different periods, or between the growth of general business activity and one’s own business.
6. It must adequately meet the many problems of logic and statistical technique which arise in its calculation.¹⁴

Series Included.—Fifty-four individual series are used in the index. The classification of these into seven groups and the number of individual series in each group are shown in Table 142.

Removal of Components Other than Cyclical.—*Calendar variation:* Some of the series are used on a plain monthly basis, others are changed

TABLE 142
MAJOR GROUPS AND SUBGROUPS OF SERIES INCLUDED IN THE
“BABSONCHART OF BUSINESS ACTIVITY” WITH BASE AGGREGATES
AND PERCENTAGE IMPORTANCE OF EACH GROUP

GROUPS AND SUBGROUPS OF SERIES	AVERAGE VALUE* ADDED BY MANU- FACTURE 1923-27 (000,000 omitted)	PERCENTAGE WEIGHTING OF GROUPS AND SUBGROUPS
Grand total (54 series).....	\$29,068	100.00
Manufactures (29 series).....	16,963	58.4
Foodstuffs (5 series).....	2,055	12.1
Textiles (8 series).....	3,445	20.3
Rubber consumption (1 series).....	623	3.7
Automobile manufactures (2 series).....	2,874	16.9
Coal and oil products (3 series).....	983	5.8
Iron and steel (2 series).....	2,660	15.7
Paper production (1 series).....	720	4.3
Printing and publishing (3 series).....	1,700	10.0
Portland cement production (1 series).....	409	2.4
Boot and shoe production (1 series).....	839	4.9
Tobacco (2 series).....	655	3.9
Mining (8 series).....	3,580	12.3
Agricultural marketings (12 series).....	644	2.2
Building and construction (1 series).....	2,799	9.6
Electric power production (1 series).....	1,018	3.5
Railroad freight revenue ton miles (1 series).....	3,454	11.9
Foreign trade (2 series).....	610	2.1

* Agricultural marketings represent only the marketings phase of agriculture computed as 10 per cent of the “cash value” of crops; value added by building and construction was taken at 45 per cent of the value of contracts awarded; foreign trade was made to represent only the direct “value added” by the handling of imports and exports, estimated as 7 per cent of value. All manufacturers’ items were raised by a constant to give that major group the “value added” importance derived from the Census of Manufactures for the sum of the groupings represented; these amounted to 64 per cent of total manufactures. Quoted from *Technical Description of the Babsonchart*.

¹⁴ From a special release of Babson’s Statistical Organization, entitled *Technical Description of the Babsonchart*. The description which follows is taken from the same source, supplemented by information supplied directly by Mr. H. C. Baldwin. All of this material is reproduced through the courtesy of Babson’s Statistical Organization.

to an average daily basis for the month. "In those cases where industries did not work 365 days a year and where the holidays observed could be ascertained with a fair degree of accuracy, adjustments were made for the working day irregularities by compiling the series on the basis of average daily output."

Price factor: Only two of the series are expressed in dollars—building and construction, and foreign trade. Both series are adjusted for changes in the general price level.

Seasonal: A seasonal pattern is determined for each series in the index by a modification of the moving-average method explained in chapter XXIII. A twelve-month centered moving average is computed and each monthly item of the original data is divided by the corresponding value of the moving average. The quotients expressed as per cents are arranged in order of size for each month separately for the most recent seven years. Thus the quotients for January of the years 1932–38, inclusive, are arranged in order as a step in obtaining the seasonal corrector for January, 1939. The average of the middle three of these ordered quotients becomes the seasonal corrector for January. The same procedure is followed month by month. The seasonal variation is removed by dividing each item of the original data, either monthly totals or daily averages, by the seasonal corrector for that month.

Combining the Series.—None of the series has an inverted cyclical movement; consequently this step in the adjustment is omitted in preparing the Babson Index.

Monthly basis: The removal of the seasonal variations leaves a set of corrected monthly series containing trend and cyclical fluctuation. Each of the series is then multiplied by the appropriate factor.

Lag: The cyclical movements of all but four of the series were found to be essentially simultaneous. These four—building, silk imports, rubber imports, and cotton takings—were shown by tests to lead the other series in their cyclical movements. Each of the four series is adjusted (lagged) by the computation of a moving average which is used instead of the original data. Thus a three-month moving average fitted to January, February, and March would center on February, and this average would be used in computing the March business index. Currently, the silk and rubber imports and cotton takings are replaced by silk deliveries, rubber consumption, and cotton consumption.

Varying amplitude: No provision is made for equalizing the effect of high- and low-amplitude cyclical fluctuations. The adverse effects of this situation are largely offset by the inclusion of a large number of series with weights so distributed that no one series has sufficient importance in the combined index to permit the pattern of its cyclical fluctuations to determine the combined pattern.

Weights: The relative importance of the several groups in the final index is shown by the per cents in the last column of Table 142. The per cents, however, are not used directly in the computation of the index which is constructed as a weighted relative of aggregates. The details of the weighting process are described in the next section in connection with the method of constructing the final index.

In most cases, the "Average Value Added" figures in the table have been adjusted so that the amount ascribed to a particular commodity is not the actual "value added" by that commodity but a figure which will give to the commodity its proper importance as a representative of the manufacturing group in which it appears. When this procedure is applied to all of the commodities the result is intended to be representative of all manufacturing. At the next stage the non-manufacturing items are adjusted so that all of the major groups have proper importance in the total index, thus 58 per cent is manufactures, 12 per cent mining, etc. The most important subgroups under manufactures are textiles, automobiles, iron and steel, and foodstuffs.

Form of the final index: The business index is constructed by the use of the weighted aggregative formula explained in chapter XIX.¹⁵ The base period is the years 1923, 1925, and 1927 and the base aggregate, $\Sigma(q_0 p_0)$, means the average "Value Added by Manufacture" in the base years for the series included in the index as described in the footnote to Table 142. In the table the value of $\Sigma(q_0 p_0)$ is \$29,068,000,000.

To obtain p_0 for each series for use in the numerator, $\Sigma(q_k p_0)$, the average "value added" in the base period, $(q_0 p_0)$, for each series is divided by the average base quantity, q_0 , for that series. Thus the average "value added" of electric power production, \$1,018,000,000, is divided by the average production in the years 1923, 1925, and 1927, 67,248,000,000 K.W.H.¹⁶ to obtain the value added per K.W.H., 1.5

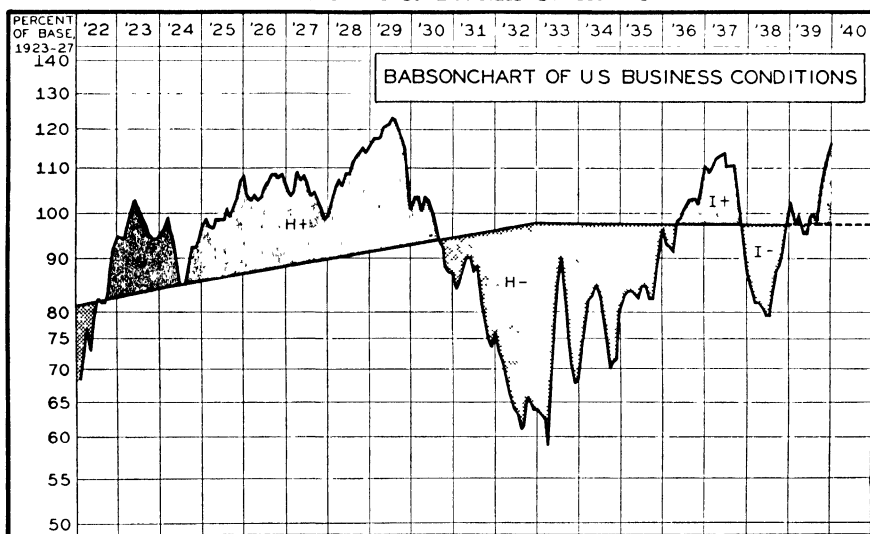
¹⁵ For quantities, the formula takes the form: $I = \frac{\Sigma(q_k p_0)}{\Sigma(q_0 p_0)}$

¹⁶ Computed from *Survey of Current Business* (Annual Supplement, 1932), pp. 142-43.

cents. This is the base "price" (p_o) for electric power production. The quantity of electricity produced in any current month is, in terms of "value added," on a monthly basis. The electricity produced in any one month is the q_k . This is then multiplied by p_o to obtain the "Value Added by Manufacture of Electric Current" in that month.

After similar computations have been completed for all of the series the results are totaled to obtain the aggregate value $\Sigma(q_k p_o)$, for the current month, and finally the quotient, $\Sigma(q_k p_o) \div \Sigma(q_o p_o)$, gives the index for the current month. The fact that Table 142 is prepared on an annual basis comes from the usual convenience of keeping census "value added" data in mind on the same basis as the census presents them.

FIGURE 99
BABSON CHART OF BUSINESS CONDITIONS



Reproduced from the *Babsonchart*, Feb. 12, 1940, with the permission of Babson's Statistical Organization.

Trend: The trends have not been removed from the individual series; hence the index described in the preceding paragraph and presented in Figure 99 contains trend and cycle. The final step consists of fitting a trend to the composite curve as shown by the X-Y line on the chart. The method of fitting is simple. The average of the monthly values of the index for each completed cycle is taken as the value of the X-Y line at the middle month of the cycle. The average points are joined from cycle to cycle to give the complete line as

shown on the chart. The position of the line is always subject to later correction for the current incomplete cycle as indicated by the dotted line extended into area J+.

The contour curve (Figure 99) is the Babson Index of the U. S. Physical Volume of Business. Tests have shown this monthly business curve to be a good representation of the nation's commercial and industrial *real* income or progress (in terms of goods and services rather than money). *The average of 1923-27 equals 100.*

Every successive business cycle has had a higher average level; the X-Y "*normal line*" connects these averages. Thus the estimated normal is clearly shown. The *areas* above the X-Y Line of Normal Growth represent periods of over-expansion in business; the *areas* below the X-Y line are periods of readjustment. An overexpansion area tends to be followed by a depression area of approximately equal size though varying shape—based on the theory that (when time and intensity are multiplied to form an area) "action and reaction" are equal.

The *ratio scales* automatically give proportional movements in the curve. Equal vertical distances between any points represent equal percentage changes.¹⁷

LOCAL INDEXES

Purpose

National indexes are constructed from data representing the entire country; consequently they give a composite view of the average situation but no information concerning differences that exist from one part of the country to another. Specifically, in the Annalist Index cotton consumption may be 10 per cent higher this year than last but that increase may be the net result of a decline in mill order: in the New England area and an increase in southern areas. This example is typical of the kind of local variations that may underlie any series representing the whole country. When a variation occurs in one series, it is usually caused by some unusual local condition that will exercise a like effect on other series in that area. Consequently the entire level of business in a given area may be considerably higher or lower than that indicated by a national index.

In statistical terms this circumstance is merely an expression of the presence of significant place dispersion in the national average. This is not a situation, however, which can be corrected by some change in the method of constructing national indexes. Most series of data cannot be separated into parts representing local areas; hence the conceivably

¹⁷ Taken with slight modification from the explanation of the Babsonchart in Babson's *Reports of January 8, 1940.*

alternative of constructing a national index as an average or aggregate of a large number of local indexes is not feasible.

The solution of the problem is the construction of local indexes to supplement national indexes. It must be understood that local indexes are not substitutes for national indexes, but merely auxiliaries developed to assist individual business communities in the conduct of their affairs. In accord with this point of view Babson, Dun and Bradstreet, Standard Statistics, and others regularly publish indicators of business conditions in local areas as supplements to the presentation of national conditions.

A local indicator of this sort is usually based on a single series such as bank debits, bank clearings, business failures, the ratio of steel production to capacity, or department-store sales. Such an indicator serves fairly well for comparing one area with another or with the whole country, but is less comprehensive than a local composite index which includes various lines of business activity in the area.

Advantages

The advantages of a composite local index have not always been well understood. A local index can be planned to take account of the importance of particular industries in the area and reduce the importance of others which, while significant nationally, may play only a secondary role or not even be represented in the business structure of a particular locality.

The differences in timing and intensity of seasonal movements as between localities are averaged in a national index and tend to cancel each other. In a local index the seasonal variation as it occurs in that locality is measured and adjusted for each series. The result will be a precise indicator of business conditions in that locality.

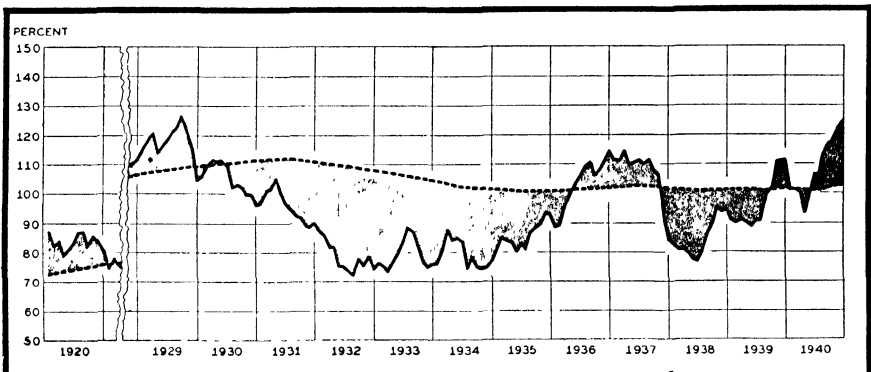
Individual concerns may wish to compare their operations with the advances and recessions of business in general. Comparisons of certain industries or lines of trade with general business may also be desired. For these purposes a local index of general business furnishes a better standard of comparison than does a national index, for the latter cannot be other than an average of good, bad, and indifferent conditions in various sections of the country at the same time. Local indexes can be interpreted in terms of local situations familiar to both maker and user of the index, and can be compared readily with national indexes.

Examples

The construction of composite local indexes has been neglected in the past for two major reasons: (1) the lack of suitable series of local data and (2) a general lack of understanding among business men of the usefulness of local indexes. A few of these have been published in recent years among which the following may be mentioned. "Index of Business Activity in Pittsburgh," published monthly by the Bureau of Business Research of the University of Pittsburgh; "Index of Business Activity in the Philadelphia Area," published monthly in the *Bulletin of the Philadelphia Federal Reserve Bank*; "Index of Business Activity in Toledo," published monthly by the Bureau of Business Research of the University of Toledo; "Index of Business Activity in Buffalo," published monthly by the Bureau of Business and Social Research of the University of Buffalo.

The Buffalo Index of Business Activity.—The methods used in the construction of a local index are not materially different from those used in constructing a national index. Special attention, however, is required in dealing with certain problems, such as the difficulty of securing continuous data for a period of years and the question of the importance locally of a given series of data. For example, it was not desirable to include any series concerning the production of textiles in the Buffalo Index because that field of production is almost wholly absent from the western New York area. Flour milling on the other hand was made unusually important because it ranks third in value

FIGURE 100
BUFFALO INDEX OF BUSINESS ACTIVITY



Reproduced from the *Statistical Survey*, January, 1941, University of Buffalo Bureau of Business and Social Research.

of products among western New York industries according to the 1937 Census of Manufactures.

Figure 100 is a reproduction of the Buffalo Index. The principal difference in the method of construction between this index and other national or local indexes is the use of moving trend. This type of trend was explained in chapter XXIV, pages 639-46. It leads to the mildly fluctuating reference or average line appearing on the chart. This line is a composite of individual trends removed at an earlier stage of the analysis from each of the eight series included in the index. These eight series with the weight assigned to each in the construction of the composite cycle and the composite trend are listed in Table 143.

TABLE 143
THE EIGHT SERIES USED IN THE CONSTRUCTION OF THE
INDEX OF BUSINESS ACTIVITY IN BUFFALO WITH THE
WEIGHT ASSIGNED TO EACH *

	WEIGHT
Steel production ratio	10
Bank debits	20
Flour milling	5
Employment	20
Postal receipts	2
New automobile registrations	3
Department-store sales	24
Industrial power consumption	16
Total	100

* From the files of the Bureau of Business and Social Research, The University of Buffalo

USE OF BUSINESS INDEXES

Direct Interpretation

The obvious direct uses of the indexes described in this chapter are to show (1) the amount of expansion or contraction of business in the current month or week, (2) the relation of the current period to the corresponding period of the preceding year, (3) the relation of recent movements to the long-run situation. Interpretations of this sort are commonplace to business executives although in many cases they are not sufficiently familiar with the methods employed in the construction of the indexes they use to understand fully the meaning of the changes that occur. In spite of this circumstance the use of business indexes has been beneficial to management.

As an Aid to Forecasting

Beyond the employment of business indexes for the knowledge of current conditions which they provide there is an additional use in the field of forecasting. It is not possible within the scope of this book to develop in detail the subject of forecasting, and in any case the methods are largely non-statistical in character. Withal the part of forecasting which is related to business indexes can be stated very simply.

A forecast of future sales, production, required inventories, or other business activity depends primarily upon conditions within the forecasting concern. The preparation of such a forecast is the task of a manager who has at his command complete data concerning the firm's operations and is thoroughly familiar with long-range policies as well as any peculiarities of the business which may affect either its immediate or more remote prospects. The use of such factors in making a forecast is largely a matter of judgment based on experience. Inevitably, however, the question arises, "What are the prospects for business in general and what is the relation of the particular business to the general prospects?" Business indexes provide the basis for answering this question, a fact which makes them an integral part of the forecaster's equipment.

The application of a knowledge of general business conditions to production planning within a business concern is explained in some detail in the next chapter.

PROBLEMS

1. What are the advantages and disadvantages of using single series as indicators of business in general?
2. Wherein lies the advantage of a composite index of business as contrasted with single-series indicators in a situation such as that created by the defense program of 1941?
3. What additional factors are involved in the construction of a business index as compared with the method of constructing index numbers described in chapter XIX?
4. Describe the methods of testing series for the existence of lag.
- 5
 - a) What are the reasons for equalizing amplitudes of the cyclical fluctuations of the individual series in preparing a business index?
 - b) What are the methods of equalizing amplitudes?
6. Explain the difference between a cyclical index and a trend-cycle index.

7. The following series represent various phases of retail distribution:

YEAR	ANNUAL TOTALS (millions of dollars)					INDEXES		
	(1) Money Orders Issued	(2) Sales of Restaurant Chains	(3) Mail- Order Sales	(4) Cost of Magazine Advertising	(5) Cost of Radio Facilities	(6) Dept. Store Sales (1923-25 = 100)	(7) Variety Chain- Store Sales (1929-31 = 100)	(8) Grocery Chain Sales (1929-31 = 100)
1932	386.6	44.0	460.1	110.7	39.1	69	80.8	85.7
1933	419.7	38.9	477.3	97.9	31.5	67	82.5	80.3
1934	411.2	42.0	595.7	116.5	42.6	75	90.5	83.3
1935	427.5	40.7	718.5	122.8	49.3	79	91.5	89.6
1936	463.1	42.4	894.2	143.6	59.4	88	99.5	94.4
1937	494.8	44.0	1,007.1	165.3	69.7	92	102.0	95.7
1938	466.0	39.4	926.8	140.4	71.7	85	98.0	94.1
1939	464.2	39.6	1,147.4	151.0	83.1	90	102.0	102.5

- a) Assuming that these series contain only cyclical fluctuations and that problems of lag, varying amplitude, and weighting can be neglected, prepare a composite index of retail trade based on these eight series.
- b) To what extent are the assumptions made in (a) justified? Discuss in detail.

8. The monthly prices of steel scrap from 1929 to 1940 and the monthly shipments of finished steel by the United States Steel Corporation are as follows:

MONTH	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938	1939	1940
PRICE STEEL SCRAP—Dollars per Hundred Gross Tons*												
January ...	1,525	1,269	1,022	750	525	1,050	1,180	1,338	1,806	1,300	1,385	1,638
February ...	1,588	1,331	1,006	716	525	1,100	1,125	1,419	1,944	1,269	1,406	1,575
March	1,556	1,319	1,000	713	525	1,213	1,050	1,475	2,085	1,215	1,425	1,569
April	1,595	1,300	981	700	600	1,175	985	1,434	2,056	1,138	1,338	1,533
May	1,538	1,250	888	640	845	1,095	1,006	1,288	1,738	1,095	1,280	1,688
June	1,494	1,206	875	569	891	975	997	1,285	1,595	1,038	1,356	1,819
July	1,475	1,200	875	488	1,041	955	1,035	1,338	1,763	1,200	1,356	1,735
August ...	1,506	1,213	838	575	1,045	919	1,238	1,519	1,970	1,375	1,388	1,803
September ...	1,513	1,250	820	625	984	850	1,250	1,615	1,756	1,350	1,622	1,922
October ...	1,430	1,138	800	600	933	875	1,250	1,625	1,469	1,288	1,905	1,975
November ...	1,313	1,013	800	593	856	925	1,300	1,650	1,250	1,420	1,766	2,006
December ...	1,250	1,000	780	525	894	1,031	1,335	1,715	1,238	1,375	1,656	2,060
SHIPMENTS OF FINISHED STEEL (Thousands of Short Tons)												
January ...	1,365	1,218	879	465	313	366	587	795	1,264	570	871	1,146
February ...	1,388	1,262	835	449	302	426	643	747	1,253	522	747	1,009
March	1,606	1,367	993	422	279	650	733	864	1,563	627	845	932
April	1,617	1,310	957	430	366	710	650	1,081	1,485	551	772	908
May	1,702	1,326	837	370	498	823	659	1,087	1,443	510	796	1,084
June	1,529	1,083	717	356	663	1,086	636	978	1,405	525	808	1,210
July	1,480	1,041	652	295	772	407	603	1,050	1,315	485	745	1,297
August ...	1,500	1,044	626	316	735	414	687	1,020	1,226	616	886	1,456
September ...	1,263	954	532	341	634	405	676	1,061	1,161	636	1,087	1,393
October ...	1,333	861	520	337	633	375	756	1,109	876	730	1,346	1,572
November ...	1,110	740	474	299	473	401	752	974	649	749	1,406	1,425
December ..	932	636	383	250	656	460	730	1,179	540	766	1,444	1,545

* This unit permits the use of the same scale for both curves on the diagram.

Plot the two series as given. Since neither one contains a discernible seasonal pattern during this period the rhythmic movements can be taken as cyclical fluctuations. Determine from the graph whether there is any lag between the fluctuations of the two series.

9. Select one of the indexes from the references at the end of the chapter. Write an explanation of the statistical methods and devices employed in preparing the index. Place special emphasis on any interesting features of the index.

REFERENCES

GENERAL

BRATT, E. C., *Business Cycles and Forecasting*. Chicago: Richard D. Irwin, Inc., 1937.

Chapter XV contains an appraisal of individual measures of business.

RIGGLEMAN, JOHN R., *The Problem of Analyzing Local Business Conditions*. Los Angeles: Eberle and Riggleman, 1927.

"The Reliability of Various Statistical Series as City and Regional Business Indexes," *Journal of the American Statistical Association*, Vol. XXIX (March, 1934).

NATIONAL INDEXES

Industrial Production and Trade *Barron's Weekly*

PERSONS, WARREN M., "Gauging Business Activity, Barron's New Weekly Business Index Explained and Compared with Yardsticks Heretofore Available," *Barron's Weekly* (Barron's Publishing Company, January 18, 1937).

Business Activity *Business Week*

"The Index of Business Activity and How to Use It," *Business Week* (September 17, 1938). New York: McGraw-Hill Publishing Co., Inc.

American Business Activity *Cleveland Trust Company Business Bulletin*

AYERS, LEONARD P., "American Business Activity Since 1790," Supplement to the Cleveland Trust Company Business Bulletin (14th ed., October, 1940).

General Economic Conditions *The Review of Economic Statistics*

PERSONS, WARREN M., "An Index of General Business Conditions," *The Review of Economic Statistics*, Vol. I (April, 1919). Cambridge: The Harvard University Committee on Economic Research.

A description of the method by which the original form of the Harvard Index was computed

FRICKEY, EDWIN, "Revision of the Index of General Business Conditions," *The Review of Economic Statistics*, Vol. XIV (May, 1932). Cambridge: Harvard Economic Society.

A description of the present form of the Index.

Business Activity *Sunday New York Times*
 ELLSWORTH, D. W., *The New York Times Weekly Index of Business Activity*. New York: The New York Times, 1933.

Production and Trade *Monthly Review of Credit and Business Conditions*,
Second Federal Reserve District
 JOHNSON, NORRIS O., "New Indexes of Production and Trade," *Journal of the American Statistical Association*, Vol. XXXIII (June, 1938).

LOCAL INDEXES

Productive Activity in Manufacturing in *Business Review of the Federal*
 the Philadelphia Federal Reserve District *Reserve Bank of Philadelphia*
 "A Description of the Construction of an Index of Productive Activity in Manufacturing in the Philadelphia Federal Reserve District," a Supplement to *The Business Review* of the Federal Reserve Bank of Philadelphia, Philadelphia, January, 1932.

Business Activity in the Pittsburgh District *Pittsburgh Business Review*
 FRITZ, WILBERT G., and VEENSTRA, THEODORE A., "Thirteen Years of Business Activity in the Pittsburgh District," *Pittsburgh Business Review*, Vol. II (March, 1932). Bureau of Business Research, University of Pittsburgh.

Business Activity in Buffalo *Statistical Survey*
 "Index of Business Activity in Buffalo," *Statistical Survey*, Vol. XIII (Supplement April, 1938). Bureau of Business and Social Research, University of Buffalo.

DAVENPORT, DONALD, and SCOTT, FRANCES V., *An Index to Business Indices*. Chicago: Richard D. Irwin, Inc., 1937.

The comprehensive list of business indexes given in pages 113-24 can be used to supplement those mentioned above.

CHAPTER XXVI

INTERNAL APPLICATION OF TIME-SERIES ANALYSIS PRODUCTION PLANNING

INTRODUCTION

THE preceding chapter contained an explanation of the application of time-series analysis in the construction of business indexes. This type of work might be called the external application of time-series analysis because it involves bringing together a number of series of data outside of one concern's records. Business indexes are usually constructed by associations, university bureaus, commercial agencies, or publishing companies, and a few individual business concerns (notably the American Telephone and Telegraph Company) engage in such work. In the latter case, however, the series analyzed are not taken from the concern's own records.

An entirely different application of time-series analysis is found in the routine statistical work connected with the business activity of an individual firm. The records used relate to the concern's own operations and the work is conducted as a part of the function of management. Such internal statistical work may take many different forms and serve many different purposes. Sometimes the statistical application becomes less important than the management aspect, and the work is carried on under the name of planning. Whatever the name or the emphasis, the work comes under the head of control of operations through the analysis of numerical records.

The extent to which a particular concern has need of such analysis depends upon the size of the concern, the type and diversity of its operations, the managerial methods employed, and other related factors. Obviously the analysis needed will vary from one concern to another. To explain all of the uses of numerical records within individual businesses would produce a discussion too diffuse for easy comprehension, too brief for a clear understanding of what is really involved in such work, and withal far from comprehensive because the complete scope of such internal analysis is unknown.

The purpose of this chapter is rather to explain in some detail the type of analysis found at one stage of the control process of a

particular concern. The Eastman Kodak Company of Rochester, New York, provided the material and the basic explanations which appear on succeeding pages.¹

GENERAL DISCUSSION OF PRODUCTION PLANNING

The simplest form of planning of production schedules consists in setting current output on the basis of previous month's sales adjusted to the best estimates of business trends in the immediate future. The details might be developed as follows: production last month was 10,000 units; sales during the month were planned for 9,000 units but actually amounted to 11,000 units. As a result, stocks on hand are somewhat depleted and there is every indication that demand will be sustained for at least two more months. Therefore a crude estimate might increase the production schedule for the current month to about 12,500 units.

This kind of planning may be fairly effective in a small concern manufacturing one or at most a few well-standardized products. It fails to take account of such elements of production control as maintenance of adequate stocks of finished goods without carrying excessive inventories, regularization of production to avoid fluctuations in employment, and normal seasonal variations in demand. Furthermore this crude method depends entirely too much upon individual judgment. Planning can, of course, never eliminate the judgment factor entirely, but a method which makes the maximum use of an objective formula and reduces to a minimum the area within which judgment determines the result is always to be preferred. This point is even stronger when the planning technique is being applied to the production of various parts which must be assembled into a single product. The planning of the production of radio sets requires a co-ordination of the work of making all of the separate parts so that none will be deficient at the point of assembly. A similar situation does not arise in the manufacture of steel rails.

In a plant manufacturing thousands of articles, whether assembled or not, the chance of securing smooth operation by the use of crude estimates is just about nil. Modern industry therefore relies on the use of numerical devices specially developed for planning and control

¹ All of this information was made available through the courtesy of Mr. A. H. Robinson, assistant treasurer of the Eastman Kodak Company, and all of the forms and explanations were prepared by Mr. Laurence M. Tarnow, head of the planning department.

These devices vary with the type of product and the organization of the management employing them.

TWO EXAMPLES OF PLANNING

The two examples which are explained in this chapter refer to two regular stock articles produced by the Eastman Kodak Company. Product "S" is a low-cost article produced in large quantities. Product "C" is a more expensive article produced in smaller quantities. Complete control is exercised by the planning department in both cases, but the form used for Product "C" is designed to provide a little more sensitive control at critical points.

All of the planning operations of the Eastman Kodak Company are conducted on a basis of thirteen four-week periods per year. Hence no reference will be made to the several months by name, but by number, as seventh period, ninth period, etc. All production planning is reconsidered at the close of each period; hence the process described here is repeated thirteen times per year for each of the thousands of products manufactured by the Eastman Kodak Company.

Planning the Production of Product "S"

The planning schedule is found in Table 145, but the method used in determining the schedule is based on preliminary data drawn from other sources. It is necessary to explain these preliminary data before entering upon the planning method in Table 145.

Per Cent Supply Required.—The entries in row (A), Table 145, will be explained with the aid of Table 144. The latter is prepared for a calendar year as shown by the thirteen periods listed in the stub. Column 1 shows the seasonal variation in sales of Product "S," sales in each period being expressed as a percentage of the year's sales. This seasonal distribution is based on the actual sales experience of the preceding five years if available or, in the case of new products, an estimate of the seasonal variation is made based on actual sales experience with similar products.²

Production of Product "S" is planned on the initial assumption of a minimum stock of three periods' supply. Column 2 shows the per-

² Methods of computing indexes of seasonal variation have been explained in chapter XXIII. The use of a seasonal index in the form of a percentage distribution as in column 1 is necessary for the type of computation carried out in Table 144. This index could be changed to the form explained in the earlier chapter by multiplying each periodic per cent by 13.

centage of the estimated annual sales rate³ which should be available at the beginning of each of the thirteen periods to meet this requirement. Thus $4.7 + 6.0 + 6.2$ or 16.9 per cent of the estimated annual sales rate should be in stock at the beginning of the first period, $6.0 + 6.2 + 8.7$ or 20.9 per cent at the beginning of the second period, and so on. The 14.6 per cent on hand at the beginning of the thirteenth period will supply the sales requirements for the last period of this year and the first two periods of next year.

If no other elements of planning were to be taken into account, column 2 would give the "percentage required supply" to be used in row (A) of the planning table (Table 145). The result would be employment varying in accord with the seasonal fluctuation of sales.

TABLE 144

PRELIMINARY SCHEDULING TO DETERMINE FOR EACH PERIOD THE SUPPLY REQUIRED AS A PERCENTAGE OF ESTIMATED ANNUAL SALES RATE—PRODUCT "S"

PERIOD	(1) SEASONAL SALES VARIATION (per cent)	(2) 3 PERIODS' SUPPLY AT BEGINNING OF PERIOD (per cent)	(3) PRODUCTION RATE DURING PERIOD (per cent)	(4) NORMAL STOCK AT BEGINNING OF PERIOD (per cent)	(5) NORMAL STOCK IN TERMS OF PERIODS' SUPPLY	(6) SUPPLY REQUIRED AT BEGINNING OF PERIOD (per cent)
I	4.7	16.9	7.8	31.2	4.6	35.1
II	6.0	20.9	7.9	34.3	4.4	39.0
III	6.2	24.8	7.9	36.2	4.2	42.2
IV	8.7	28.3	7.9	37.9	3.9	44.1
V	9.9	29.8	7.9	37.1	3.6	45.8
VI	9.7	32.0	7.8	35.1	3.3	45.0
VII	10.2	33.2	7.8	33.2	3.0	42.9
VIII	12.1	30.8	7.8	30.8	3.0	41.0
IX	10.9	23.8	6.0	26.5	3.6	38.6
X	7.8	17.7	7.8	21.6	4.0	32.5
XI	5.1	13.8	7.8	21.6	4.5	29.4
XII	4.8	13.4	7.8	24.3	4.8	29.4
XIII	3.9	14.6	7.8	27.3	4.7	32.1
	100.0		100.0			

This, of course, would be unsatisfactory to labor and inefficient from the production point of view. Therefore the assumption of a uniform three-period stock is used merely as a point of departure to be adjusted upward, but not downward, in introducing the next planning element, namely, the assumption that the production rate must be kept regular throughout the year. This feature is shown in column 3 in which production is equalized in all periods except the ninth. Provision is made during the ninth period to take this particular operation out

³ The method of computing the estimated annual sales rate is explained in connection with Table 146 and Figure 101.

of production for several days for repairs to equipment and to reduce the production rate somewhat to allow for vacations which are granted to all employees.

Column 4 gives the adjustment of the three-period supply necessary to provide regular employment throughout the year. This is accomplished by allowing stocks to increase during the slack season of sales to provide for heavier-than-average demand in the peak season of sales. It is adjusted, however, so that at least three periods' supply will be available at all times.

The computation of the "normal stock" should start, in most cases, at the period where the "three periods' supply" appears as the largest (33.2 in Column 2, Table 144). By adding to 33.2 the seventh-period production and deducting the seventh-period sales, the "normal stock" for period eight will be obtained ($33.2 + 7.8 - 10.2 = 30.8$ in Column 4). Using 30.8 (Column 4) and adding production of 7.8 in the eighth period and subtracting sales of 12.1 we have 26.5 (Column 4). By repeating this process for each succeeding period it will be found that when the cycle has been completed, or after adjusting for sixth-period production and sales, we return to the starting point (33.2) and thus the calculations are proved to be correct.

When the seasonal curve changes abruptly or contains more than one peak during the year, the process described in the preceding paragraph may result in a "normal stock" figure which in one or more periods is less than the required "three periods' supply." It is then necessary to make an adjustment that consists in starting to write the "normal stock" column from the period in which the preliminary "normal stock" computation shows the greatest deficiency below the required minimum "three periods' supply."

Column 5 is used for reference to determine the number of periods' sales that would be provided by the "normal stock" at various times of the year and to supply some indication of the turnover rate that can be expected. Thus the 31.2 per cent stock at the beginning of the first period would be sufficient to provide the sales volume of the first four periods ($4.7 + 6.0 + 6.2 + 8.7 = 25.6$) and leaves 5.6 per cent of the 9.9 per cent sales of the fifth period. This is, $5.6 \div 9.9 = .56$ of the fifth-period sales; hence the "normal stock" at the beginning of the first period would provide sales for 4.56 periods or, to one decimal place, for 4.6 periods.

The excess stock dwindles gradually as sales expand in the early

periods and reaches the minimum "three periods' supply" at the beginning of the seventh period. It remains at three periods' supply at the beginning of the eighth period because production during the seventh period is equal to sales during the tenth period. After the eighth period the stock expands again to a maximum of 4.8 periods' supply at the beginning of the twelfth period.⁴

Column 6 indicates the "supply required" at the beginning of any period and is computed by adding the seasonal sales variation for the previous period to the "normal stock" for the current period. This is necessary so as to provide for sales for the previous period, which are unknown at the time of making the plan for the current period. The estimated status of stocks at the beginning of the current period must be based on inventories available at the beginning of the preceding period. Thus for the first period, Normal Stock I + Sales XIII = Supply Required I, i.e., $31.2 + 3.9 = 35.1$.

The Planning Schedule.—The complete operation is presented in Table 145. The parts of this table are closely related and the processes carried on require considerable explanation to be understood by persons not regularly engaged in maintaining these records. Row (O) is brought into Table 145 from Table 146 and Figure 101 but it has seemed preferable to ask the reader to accept row (O) without explanation temporarily.

The planning steps which follow are repeated for each period; therefore a complete explanation of the data in Table 145 would involve seven separate statements. To avoid the repetition of method involved in a complete exposition, the explanation will be confined to the planning of production for the seventh period.

The plan for the seventh period must be made during the sixth period. Hence the most recent stock and sales figures are for the close of the fifth period. The terms used in Table 145 can be understood best by thinking of the planning department as the purchaser of Product "S" from the production department.

Explanation of lower half of Table 145: The preliminary information needed for planning is recorded in rows (A) to (P), Table 145. Row (A) is taken from Table 144.

Row (B) Any time that the stock on hand in row (L) was reported less than 8,000 this schedule would be marked "on low stock" and would be set aside for replanning, and production rates would be increased.

⁴ It should be noted in the preceding that expansion and contraction refer to periods' supply and not to the percentage of yearly production represented in normal stock.

- Row (C) As long as the schedule continued to be marked "on low stock" special attention would be given to delivery dates of additional stock from the production department.
- Row (D) Scheduled production is recorded each period from the upper half of the schedule. This will be explained in the next section.
- Row (E) If the production department fails to complete the amount scheduled in any period the undelivered balance is entered in this row for the following period.
- Row (F) is the sum of rows (D) and (E).
- Row (G) shows amount completed by the production department.
- Rows (H) to (L) show the stock of finished products in Rochester and in those branch sales divisions in which Product "S" is kept in stock.
- Rows (M) and (N) show the sales per period for the current year to date and the preceding year. Sales are obtained as follows (illustrated for the fifth period): stocks at the end of the fourth period plus deliveries during the fifth period minus stocks at the end of the fifth period equal sales during the fifth period. That is, $(21438 + 7000) - 23905 = 4533$.
- Row (O) is explained in Table 146.
- Row (P) is cumulative from rows (M) and (N).

Explanation of upper half of Table 145: The actual planning is carried out in the upper part of the schedule. The supply required at the beginning of any period (first column at the left) is obtained by taking the percentage supply required in row (A) times the latest available estimated annual sales rate at the end of the second preceding period, row (O). Thus for the seventh period 42.9 per cent of 67,000 = 28,743 units. The number of units available (next column) is obtained by adding the stock on hand at the end of the fifth period, row (L), the scheduled production of the sixth period, row (D), and the undelivered balance at the beginning of the sixth period, row (E), and subtracting from this sum the supply required at the beginning of the seventh period. Thus: $(23905 + 6000 + 1000) - 28743 = 2162$.

The supply available is the sensitive column which is used mainly to determine whether the previous schedule for the current period should be altered. If the sales from period to period followed the seasonal pattern exactly, the available supply would always be zero. The progress of the plan from period to period is shown in successive rows of columns 1 to 13. To see just how this works it is necessary to follow the figures from the beginning of the year. With estimated annual sales of approximately 70,000 units the production rate would stand at about 5,000 units per period throughout the year. This sched-

ule was planned at the beginning of the year when 5,000 units were ordered by the planning department for the first period and a tentative schedule of 5,000 units for each of the following four periods was set up, row 1. The figure 12,000 recorded in the sixth period shows merely the undistributed balance of an order of 20,000 units as of that time.⁵

At the beginning of the second period, row 2, the available supply had dwindled from 3,118 units to 696 units but stocks on hand were still slightly more than required. The drop in available supply from the first to the second period shows that sales during the thirteenth period of the preceding year exceeded expectations, but there is no need to increase schedules so long as supply available contains a surplus. Accordingly production for the second through the sixth periods was planned at 5,000 units per period, a confirmation of the schedule set up for the first period.

When the planning of Product "S" was taken up for the third period, row 3, it was found that sales for the first period had far exceeded expectations and that available supply showed a deficit of 5,475 units. The schedule was raised to 6,000 units for the third period and tentatively to 8,000 units for the fourth and fifth periods. This increase exhausted the 7,000 units originally scheduled for the seventh period, but no new order for 20,000 units was placed pending the outcome of this stepping-up of scheduled production.

For the fourth period, row 4, the deficit in available supply had increased to 5,703 units showing the continuation of the high sales rate, row (M), during the second period. The proposed increase of scheduled production to 8,000 units was therefore confirmed to the production department. The tentative plan for the fifth and sixth periods was retained and a new order for 20,000 units was placed. The latter was tentatively scheduled at 6,000 units during the seventh and eighth periods.

The available supply for the fifth period, row 5, showed a deficiency of 3,901 units, an indication that the increase in schedule during the fourth period had been sufficient to cope with the expansion in sales volume, but that higher schedules were still necessary. Therefore the 8,000 unit schedule was confirmed to the production department and

⁵ On this particular item the planning department places orders with the production department for 20,000 units at a time but does not schedule its orders specifically more than six periods in advance. In this case the former order will be completed during the fourth period and a new order will be placed prior to that time, allowing the required number of periods for process time for completion of the first lot scheduled from the order.

TABLE 145

FORM USED IN PLANNING THE PRODUCTION OF PRODUCT "S"
FILLED IN TO ILLUSTRATE SCHEDULING FOR THE SEVENTH
PERIOD OF 1938

	SUPPLY REQ	AVAIL	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1)	22815	3118	5000	5000	5000	5000	5000	12000							
(2)	25350	696		5000	5000	5000	5000	5000	7000						
(3)	29540	-5475			6000	8000	8000	5000	0	0					
(4)	30870	-5703				8000	8000	5000	6000	6000	8000				
(5)	32060	-3901					8000	6000	6000	6000	4500	2500			
(6)	31500	-2062						6000	6000	6000	4500	2500	0		
(7)	28743	2162							6000	5000	4000	4000	5000	14000	
(8)															
(9)															
(10)															
(11)															
(12)															
(13)															
(A) % SUPPLY REQUIRED			35.1	39.0	42.2	44.1	45.8	45.0	42.9						
(B) MINIMUM STOCK ON LOW STOCK															
(C) DELIVERY PROMISED															
(D) SCHEDULED PRODUCTION			5000	5000	6000	8000	8000	6000							
(E) BALANCE UNDELIVERED								1000							
(F) TOTAL DUE			5000	5000	6000	8000	8000	7000							
(G) DELIVERIES			5000	5000	6000	8000	7000								
(H) STOCK & TRANSIT ROCHESTER			14807	16508	17002	19144	21829								
(I) NEW YORK			1662	842	932	994	620								
(J) CHICAGO			1812	660	716	240	736								
(K) SAN FRANCISCO			784	1157	1309	1060	720								
(L) TOTAL STOCK			19065	19167	20159	21438	23905								
(M) 1938 SALES			6981	4898	5008	6721	4533								
(N) 1937 SALES			2950	2862	7180	5700	4648	6226	5670	9170	8908	5333	3632	3824	4887
(O) ESTIMATED ANNUAL SALES RATE			70000	70000	20000	70000	67000								
(P) SALES PAST 13 PERIODS			75021	77057	74885	75906	75791								
PRICE, NET .23 LIST 35 APP CARD X2622 MFGR 20000															
1938															
PRODUCT "S"										PROMISED		LOW STOCK		BACK ORDER	

the schedule for the sixth period was increased to 6,000 units. The schedule for the ninth and tenth periods was tentatively tapered off to avoid placing another order for 20,000 units until more information became available concerning the permanence of the sales bulge during the first three periods of the year.

The deficit in available supply had dropped to 2,062 units at the beginning of the sixth period, row 6, indicating that the stocks on hand were approaching the desired normal. Accordingly the tentative schedule of 6,000 units was confirmed, but no new order was placed.

Sales during the fifth period (4,533) dropped below seasonal expectations ($70,000 \times 9.9 = 6,930$), and the available supply for the seventh period changed from a deficit to an excess of 2,162 units, row 7. The tentative schedule of 6,000 units for the seventh period was confirmed so as to avoid reducing employment too rapidly. The tentative schedule for the eighth period was reduced to the normal 5,000 units and a reduced schedule of 4,000 units was set up for the ninth period to allow for vacations in the plant as originally planned. The schedule for the tenth, eleventh, and twelfth periods was tentatively set at 5,000 units each and a new order for 20,000 units was placed.

This description brings the planning of Product "S" up to the present.⁶ As soon as the sales information for the sixth period becomes available the plan for the eighth period will be made. Thus the scheduling of production of this product becomes a continuous periodic process. The two key figures in the operation are the available supply and the estimated annual sales rate, of which the latter remains to be explained.

Estimated Annual Sales Rate.—The method of obtaining the estimated annual sales rate is explained in Table 146 and Figure 101. Both the table and the chart are set up to show how the annual sales rate was estimated for the seventh period of 1938.

Column 1, Table 146, is a repetition of the estimated seasonal sales variation from column 1, Table 144. Column 2 is the actual sales of each period taken from rows (M) and (N), Table 145. Column 3 is obtained by dividing each figure of column 2 by the corresponding figure of column 1. If actual sales were influenced by nothing but seasonal variation following the exact pattern of column 1, the figures in column 3 would all be equal. Therefore the variations found in column 3 arise from the presence of something other than seasonal in the data. If an increasing trend were present, the figures near the end of the year would be greater than those at the beginning and conversely for a decreasing trend. The shape of a cyclical movement

⁶ The material for this chapter was obtained from Eastman Kodak Company in July, 1938.

TABLE 146
COMPUTATION OF SEASONALLY ADJUSTED ANNUAL SALES RATE—PRODUCT "S"

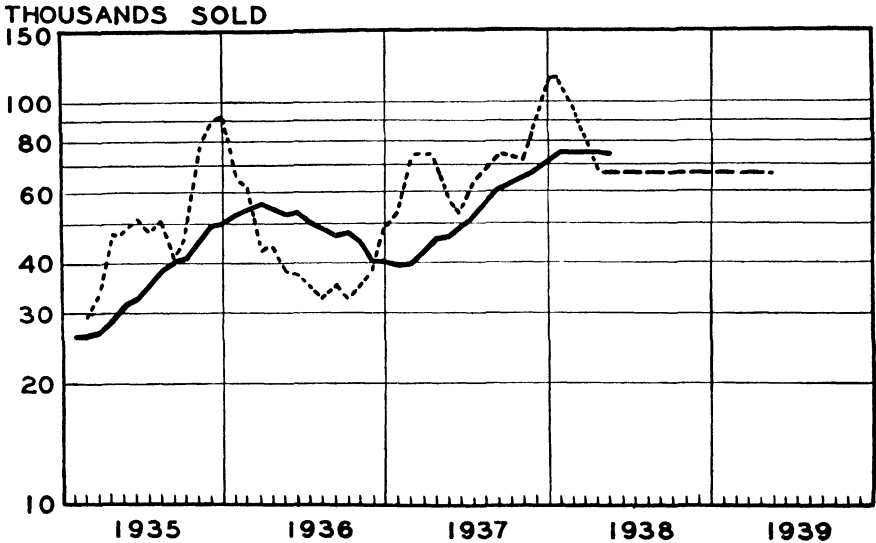
YEAR	PERIOD	(1) SEASONAL SALES VARIATION (per cent)	(2) ACTUAL PERIOD SALES	(3) ANNUAL SALES RATE ADJUSTED FOR SEASONAL	(4) 3-PERIOD AVERAGE ANNUAL SALES RATE (centered)
1937	I	4.7	2,950	63,000	
	II	6.0	2,862	48,000	75,700
	III	6.2	7,180	116,000	76,300
	IV	8.7	5,700	65,000	76,000
	V	9.9	4,648	47,000	58,600
	VI	9.7	6,226	64,000	55,600
	VII	10.2	5,670	56,000	65,300
	VIII	12.1	9,170	76,000	71,300
	IX	10.9	8,908	82,000	75,600
	X	7.8	5,333	69,000	74,000
	XI	5.1	3,632	71,000	73,300
	XII	4.8	3,824	80,000	92,000
	XIII	3.9	4,887	125,000	117,600
1938	I	4.7	6,981	148,000	118,000
	II	6.0	4,898	81,000	103,000
	III	6.2	5,008	81,000	79,600
	IV	8.7	6,721	77,000	68,000
	V	9.9	4,533	46,000	
	VI	9.7			
	VII	10.2			
	VIII	12.1			
	IX	10.9			
	X	7.8			
	XI	5.1			
	XII	4.8			
	XIII	3.9			

would likewise appear in column 3. But more important than the measurement of either of these components is the fact that short time irregularities are greatly exaggerated by converting the periodic sales into annual rates. This gives a very sensitive measure of changes from period to period resulting from other than seasonal causes. In fact the rates of column 3 have been found by experience to be too sensitive; hence column 4, a three-period moving average of the figures in column 3 has been added to the computation.

Column 4 appears as a dotted line in Figure 101. The solid line is a thirteen-period moving total of actual sales taken from row (P), Table 145. This line serves the same purpose as a moving average with a one-year period, i.e., it measures trend and cyclical movements free from seasonal variations. The cyclical movements, however, are lagged behind their actual occurrence, because the total of thirteen periods is plotted to coincide in time with the thirteenth period and

not with the seventh period as would be the case in trend-cycle analysis. This lag would be objectionable in planning the production of a product for which demand was largely dependent upon the business cycle but is of negligible importance for most of the products included by the Eastman Kodak Company under the classification of Product "S."

FIGURE 101
PLANNING CHART FOR PRODUCT "S"



Each point on solid line indicates actual sales for past 13 periods (Data from Table 145, row P). Each point on dotted line indicates average three period annual sales rate adjusted for seasonal (Data from Table 146, column 4). Broken line shows estimated sales rate for next 13 periods. *Reproduced by courtesy of Eastman Kodak Company.*

Figure 101 contains two measures of the annual sales rate of Product "S." The solid line is free from the effect of short-time irregularities. The dotted line contains all of these short-time irregularities in moderately exaggerated form. Both of them measure the general direction of trend and the shape of cyclical movements. Both are free of seasonal influences. By studying the direction of the solid line and the likely change in its direction indicated by the course followed by the dotted line, a prediction indicated by the broken line is made for the immediately ensuing periods. Thus for the seventh period of 1938 the estimated annual sales rate was placed at 67,000 units because the dotted curve has shown a declining tendency during the first five periods of the year and this short-term irregularity has already stopped the growth

of the solid curve. The prediction is based on the belief that declining sales during the sixth and seventh periods will turn the solid curve downward in the immediately succeeding periods.

This prediction of 67,000 units is transferred to Table 145, row (O), and becomes a key figure in scheduling production for the seventh period. On the chart the prediction line has been extended for a year at the 67,000 unit level. This is only tentative since the chart is kept current each period. Any time that a significant change is recorded in either the solid or the dotted curve the estimated annual production rate will be adjusted to take account of events as they occur.

Summary of Scheduling Production of Product "S."—The whole planning process is a combination of judgment and formula. Objective formulas have been used as far as possible to free the results from the effects of personal bias. Yet the planning process can never become wholly objective since the projection of the present situation into the future necessarily involves some assumptions concerning the expected direction of general business and the demand for the particular product being scheduled. The process used by the Eastman Kodak Company recognizes fully the importance of the judgment factor, but aims to minimize the possibility of distorted planning by bringing in the personal factor at several points in the operation in small doses rather than having the entire result depend upon judgment at one major stage.

Specifically judgment enters the work at three points. The first of these is the decision to maintain a minimum stock equal to three periods' supply. The second is the use of a graphic projection method to determine the estimated annual sales rate. The third and most important is the setting of the actual schedule as illustrated in the upper half of Table 145. But judgment is so closely hedged by objective measurement that there is slight chance of the planning operation breaking down as the result of failure of the personal element.

Planning the Production of Product "C"

The general principles of planning explained in connection with Product "S" are also applicable to Product "C," but the method used for the latter differs fundamentally from that previously explained. Product "C" is a more expensive article and is sold in smaller quantities than Product "S." The method of planning used provides for more precise scheduling and a greater amount of checking on schedule and sales from period to period.

Explanation of the Scheduling of Product "C," Preliminary Stage.—

The actual work is performed on the form exhibited in Table 147. Each period a new form is drawn up and the whole scheduling process is reconsidered. This is different from the form for Product "S" which is used for thirteen periods.

The information in the table which can be prepared in advance is indicated by the use of a typewriter. This particular table was used to plan production for the seventh period of 1938. The seasonal sales variation, column 2, is based on the experience of previous years. The normal production rate, column 13, expresses the initial planning element of regularizing production so as to maintain steady employment throughout the year. As in planning Product "S" so also here, the ninth period is set at a reduced rate to provide for vacations of employees. The normal stock in terms of periods' supply, column 12, is the result of computations for Product "C" similar to those explained for Product "S" in connection with Table 144. This process is not repeated since it involves no new methods, but it should be noted that the scheduling of Product "C" is based on two periods' domestic demand as a minimum supply whereas a three-period minimum supply was used for Product "S." The minimum supply is reached in the twelfth period and excess supply is built up in the other periods to meet seasonal demand without affecting the production rate.

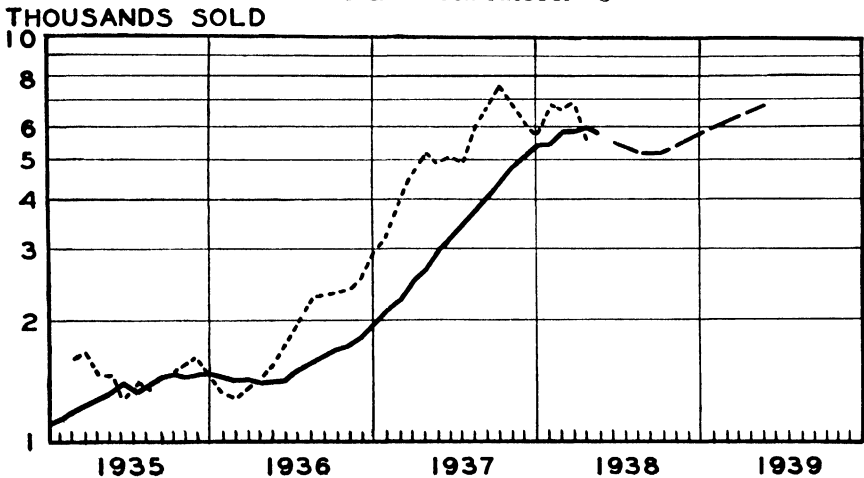
The planning of production for periods seven to twelve inclusive is carried out as soon as the actual domestic sales figure for the fifth period becomes available. This sales figure is entered in the upper left-hand corner of the table. The domestic sales for the fifth period of 1938 were 367 units, which leads to an annual seasonally adjusted rate of approximately 4,400 units, i.e., $367 \div .083 = 4,422$. The three-period average annual sales rate of 5,600 units is obtained from a table similar to Table 146 for Product "S." The computation is.

PERIOD	PERIODIC SALES	SEASONAL SALES RATE	ANNUAL SALES RATE CORRECTED FOR SEASONAL	THREE- PERIOD AVERAGE (centered)
III	406	5.8	7,000	
IV	315	5.8	5,400	5,600
V	367	8.3	4,400	

The three-period average is plotted in Figure 102 for the fourth period of 1938 (dotted line). The solid line is the moving total domestic sales for the latest thirteen periods. The figure for the fifth

period of 1938 (5,919) appears at the top of column 3, Table 147. The broken line on the chart represents the estimate of the path which the total annual sales will follow for the next thirteen periods. The prediction for the next four periods is based on the fact that during the past seven periods the dotted line has exhibited a reversal of its strong upward tendency and the fact that the actual sales total had leveled off in preceding periods and in the fifth period declined. The path followed subsequently by the broken line represents the belief of the planning department that the decline in business would terminate after the middle of 1938, and would be followed by a mild increase. The values of the broken line are read from Figure 102 to column 1, Table 147.

FIGURE 102
PLANNING CHART FOR PRODUCT "C"



Each point on solid line indicates actual sales for past 13 periods. Each point on dotted line indicates average three period annual sales rate adjusted for seasonal. Broken line shows estimated sales rate for next 13 periods. *Reproduced by courtesy of Eastman Kodak Company.*

This step is the equivalent for Product "C" of the determination for Product "S" of row (O), Table 145, from Figure 101. However, the next step is different. The estimated annual domestic sales rate for each period is multiplied by the percentage seasonal sales variation to obtain an estimated domestic sales figure for each period, i.e., column 1 $\times$ column 2 = column 3. Estimated domestic sales for the next thirteen periods total 5,770 units, slightly less than the 5,919 units actually sold during the past thirteen periods.

The foreign sales area for Product "C" is divided into four parts designated A, B, C, and D. The sales in parts A, C, and D are small and do not follow a recognizable seasonal pattern; hence representatives in the countries included in these divisions simply submit yearly estimates of the number of units they expect to order during the next thirteen periods as a whole. These combined estimates for territories A, C, and D amounted respectively to 650, 65, and 195 units in round numbers. One-thirteenth of these totals was allocated to each period for each territory as shown in columns 4, 6, and 7. Sales in territory B are larger and better adapted to control by individual periods; hence distributors in this territory submit in each period estimates of their periodic orders for the ensuing thirteen periods. These estimates are combined to give the figures shown in column 5 of the table.

Column 8 gives the expected domestic sales and foreign shipments combined for the thirteen periods. The overall estimate of 9,835 units is more than 10 per cent less than 11,028 units, the actual sales and shipments of the preceding thirteen periods.

Final Stage.—At the outset of the final stage the stock on hand at the beginning of the sixth period (1,557 units) is entered in column 9 from the stock record and the previous schedule of 800 units in the sixth period is entered in column 10. The previous schedule is also entered in column 14.

The next step is to determine the normal stock at the beginning of the thirteenth period, as shown near the bottom of the table. The normal domestic stock at the beginning of the thirteenth period should be 2.6 periods' supply according to column 12. This would mean 850 units for the thirteenth period, 270 units for the first period, and $.6 \times 350 = 210$ units for the second period, a total of 1,330 units as the normal domestic supply at the beginning of the thirteenth period. The estimated foreign shipments during the thirteenth period, 50, 140, 5, and 15 units, respectively, for territories A, B, C, and D are added to the domestic normal stock requirements of 1,330 units to give a total normal stock of 1,540 units at the beginning of the thirteenth period. In setting the normal supply one period's foreign shipments were included whereas 2.6 periods' domestic sales were included. This distinction is based on past experience showing that the buffer stock to provide for short-run changes in demand must be carried by the Eastman Kodak Company for domestic distributors but because of the

greater time required for transportation must be carried by foreign distributors.

The production required for the sixth to the twelfth periods inclusive will be the estimated sales and shipments for the seven periods plus the normal stock required at the beginning of the thirteenth period minus the stock on hand at the beginning of the sixth period. The computation is $(5,090 + 1,540) - 1,557 = 5,073$. This figure is entered at the bottom of column 10 and, rounded off to 5,100 units, becomes the amount to be scheduled during the sixth to twelfth periods, inclusive.

The guide for actual scheduling is at the bottom of column 8. The average production per period at the current estimated sales rate is $9,835 \div 13 = 757$ units, or approximately 750 units per period. On the other hand the 5,100 units estimated as the requirement for the sixth to twelfth periods, inclusive, is below that average. Eight hundred units have been scheduled for the sixth period leaving 4,300 units to be scheduled during the remaining six periods. It would appear that four of the six periods should be scheduled at 700 units, and further that the cut to 700 units should be made immediately since the stock on hand at the beginning of the seventh period will be somewhat greater than 2.8 periods' domestic supply plus foreign shipments during the seventh period, i.e., $1,602 > 485 + 425 + (.8 \times 380) + 50 + 290 + 5 + 15$. However, this reduction is unnecessary since less production during the ninth period has been planned to provide for vacations. Hence scheduled production is reduced to 550 units during the ninth period and the other periods are held at 750 units as shown in column 10. Comparison of the new schedule with the previous schedule shows that production for the tenth and eleventh periods has been tentatively stepped up 50 units. Otherwise this sheet is a confirmation of the previous one.

Two columns of the table remain to be explained. Column 9 serves immediately as a check on the computations. The estimated stock at the beginning of any period is the stock on hand at the beginning of the preceding period plus production for the preceding period minus estimated sales for the preceding period. Thus,

$$\begin{aligned} 1,557 + 800 - 755 &= 1,602 \\ 1,602 + 750 - 845 &= 1,507 \\ 1,507 + 750 - 920 &= 1,337 \\ &\text{etc.} \end{aligned}$$

The stock on hand at the beginning of the thirteenth period, column 9, must be equal to the normal stock at the beginning of the thirteenth period, bottom of column 8. The difference, $1,567 - 1,540 = 27$, is due to rounding off production from 5,073 to 5,100 units. Column 9 is also used for comparative purposes when the actual stock record becomes available from period to period.

As previously explained, the planning department of Eastman Kodak Company stands in the relation of purchaser from the production department. Column 11 simply shows how the balance of 5,500 units of Product "C" in process will be depleted by production during successive periods. At the beginning of the thirteenth period only 400 units of the orders in process will remain to be filled. It is not necessary, however, for the planning department to assume the responsibility for placing a new order at this time since the orders in process are sufficient to carry through the twelfth period according to present schedule. If no significant changes appear when the scheduling is repeated for the eighth period, presumably a new order will be placed for delivery during the thirteenth and following periods. The placing of these orders in advance by the planning department is necessary so that the production department will have sufficient time to obtain materials, process the goods, and have an opportunity to do its own internal planning.

The whole system can be, and sometimes of necessity is, placed on a hand-to-mouth basis by failure of the planning department to place orders in advance of schedule requirements. Smooth operation, of course, calls for avoiding this type of delay as far as possible. It can therefore be assumed that the planning department will fail to place orders in advance only because some emergency circumstance has arisen in connection with the particular product or products thus delayed.

SUMMARY

The two schedules explained in this chapter are representative of two types of planning routine followed by the planning department of the Eastman Kodak Company. There are, of course, many variations necessary to adapt the forms to the requirements of particular products, but the general principle followed is either the percentage type of calculation explained for Product "S" or the actual unit calculation explained for Product "C."

The feature of the scheduling process which marks it as applied statistical technique is the constant interweaving of conditional judgments and arithmetic computations. The scheduling process can never be exact, since it involves prediction of future sales, but the device of using exact methods as far as possible and spreading the judgment factor as much as possible reduces the chance of "missing the market," aids in minimizing excess stocks without jeopardizing inventory requirements, and contributes greatly to the maintenance of regular employment.

A further factor in accomplishing the goal set forth in the preceding statement is the continuous nature of the planning process. Each period the schedule for each product is reconsidered by the planning department. For lower-cost articles such as Product "S" the reconsideration consists mainly of a check of the existing schedule to see that nothing has occurred to disturb the established routine. For higher-cost articles such as Product "C" the whole planning routine is carried out anew each period. As a further control weekly reports on Product "C" are received by the planning department from the sales department and the stockroom. If at any time these interim reports should indicate that a deviation from the assumptions underlying the plan had occurred, the whole schedule would immediately be reconsidered and altered according to the new conditions.

The two examples presented in this chapter illustrate two distinct methods of production planning used by the Eastman Kodak Company, but they give no indication of the scope of the whole planning process. The planning department alone deals with about four thousand articles each four-week period and in addition maintains weekly control over the more important articles manufactured by the company. But this central control is only one of the planning operations found in the organization. The production department must carry on an entirely different type of scheduling designed to maintain an even flow of parts and assemblies to provide the finished products required for the completion of the schedule issued by the central planning department.

Close contact with the sales department is maintained by the planning people to co-ordinate sales effort with the various factors affecting demand. Similar contact is maintained with the advertising department. Development and engineering activities are continually checked to determine possible influences of new products or improvements on existing products. Again, stocks of finished products in the several

domestic branches and in branches located in foreign countries must be maintained through proper planning by the respective managers. Thus it becomes apparent that planning, which occupies a major position in the central office, is equally important in other phases of a concern's operations.

A question naturally arises as to the transferability of the methods explained in this chapter to other types of manufacturing operation. Each concern has problems of production planning peculiar to the kind of goods manufactured, the firm's type of organization, the nature of demand for products, and similar affecting circumstances, but the principles of control and the methods of achieving that control as set forth in the chapter are applicable in general. This in fact is the principal reason for including the chapter in a statistical text. If potential statisticians possess an advance knowledge of the problems of production planning and the general principles employed in their solution, such knowledge should form a sound foundation upon which to build detailed knowledge of the planning methods used by a particular concern. It should provide the background necessary to establish planning techniques in a new concern or for that matter to improve those employed by an old concern.

From this point of view the value of the chapter lies in the attempt to show beginners in the subject of statistics an important application of the subject within an individual concern. Beyond the direct knowledge of planning methods contained, the chief lesson to be conveyed by this material is the extent to which general statistical methods must be modified to meet the needs of particular applied problems and the degree to which numerical exactness is interspersed with judgment based on experience in arriving at results. It cannot be stressed too much that applied statistics involves experience as the basis for interpretation of results as much as knowledge of techniques for analysis.

PROBLEMS

1. The seasonal indexes in this chapter are percentage distributions. How does this form of expressing seasonal differ from that developed in chapter XXIII?
2. State the key operations in planning the production of Product "S."
3. State the key operations in planning the production of Product "C."

4. A subsequent report shows that the stock of Product "S" at the beginning of the seventh period in 1938 was 24,649 units. Carry out the complete planning of the production of "S" for the *eighth* period.
5. The sales for Product "C" for the *sixth* period of 1938 were 490 units. Draw up a new form and carry out the complete planning operations for the *eighth* period, including the graph.

CHAPTER XXVII

CORRELATION

INTRODUCTION

THE central purpose of statistical analysis is the development of additional and more effective methods of comparison. A measure of correlation is a major addition to this analytical equipment. The distinguishing feature of the new technique can be explained by a simple example. Suppose that we were to take from the office of a school of business administration the grades of all sophomores in two required courses, statistics and marketing. From these two lists of grades of identical students, averages and dispersions could be computed in the usual way. These results might show that the average grade was five percentage points lower in statistics than in marketing, but that the marketing grades were more uniform than those in statistics. These measures provide comparative summary information, but they give no indication of the relative performance of individual students in the two courses. Studies of the relation between individual pairs of grades can be made by the methods of correlation.

As stated in chapter III, statistical techniques are designed for the analysis of mass phenomena. Accordingly, interest in the lists of grades is not centered in the comparison of particular pairs of grades but in the broader question of the prevailing nature and extent of the association between grades by pairs in the two courses for the entire class. That is, do students who have high grades in one course tend to have high grades in the other? Are the same students likely to receive low grades in both courses? Is the association inverse, students who have high grades in one course being low in the other and vice versa? Or is there no tendency for good grades in one course to be associated with either good or poor grades in the other?

The purpose of correlation analysis, therefore, is quite distinct from that served by the measures developed in preceding chapters in dealing with data classified according to a single variable. On the basis of the illustration of grades a definition of correlation can be formulated. *When a set of items is recorded with respect to the values of two distinct variables and it is found that corresponding values of the*

two variables tend to be associated, either directly or inversely, the two variables are said to be correlated. This definition does not imply any causal relation between the two variables. In the illustration of grades there is no suggestion that grades in marketing are the cause of grades in statistics or the reverse. The comparison rests on the assumption that the group is homogeneous and that the mental capacity of a given student is the same in each of his several courses. Therefore, comparative achievement is related closely enough to warrant joint study. If the two variables were grades of identical students in *sequential* courses, Statistics I and Statistics II, for example, it is quite probable that the association would be the expression of a cause-and-effect relation.

In dealing with a cause-and-effect relation, the cause is considered to be the independent variable and is denoted by X . The effect is considered to be the dependent variable and is denoted by Y . When no such relation is present, the X -variable will be taken as the one to which the Y -variable is to be compared. It is not always obvious which variable is dependent upon the other, and sometimes it is desirable to make comparisons both ways. In such cases it makes little difference which variable is designated as X and which as Y .

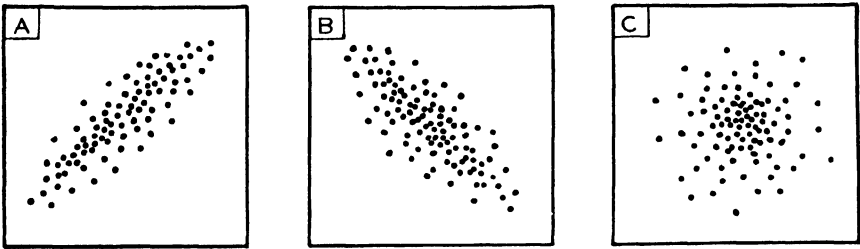
Students might easily infer from this discussion that pairs of corresponding values of any two variables comprise legitimate material for correlation analysis. While it is true that the comparison of any two variables can be made according to the definition, it does not follow that all such comparisons lead to fruitful analyses. There is a moderate negative correlation between the percentage of the population in the several states of the United States that are members of Masonic societies and the average horsepower produced per electric power plant in the same states. But it would be difficult to discover any valid reason for making this comparison and equally difficult to attach any meaning to the result.

This example shows that the initial question as to what variables can be correlated must be answered by the investigator. If a cause-and-effect relation exists, correlation analysis can always be employed. Beyond the realm of cause and effect there are three circumstances in which correlation is justified, (1) when the correlated variables are both dependent upon a third underlying variable, (2) when a close connection exists between the correlated variables, (3) when a definite reason for correlation resides in the purpose of a particular study.

SCATTERGRAM

The simplest method of studying correlation is by the use of a two-dimensional graph known as a "scattergram." The horizontal or X-axis represents the independent variable, and the vertical or Y-axis the dependent variable. That is, regardless of whether or not a definite causal relation exists between the two variables, the X-axis is used for the variable to which the other, Y, is being compared.

FIGURE 103
THREE SCATTERGRAM PATTERNS



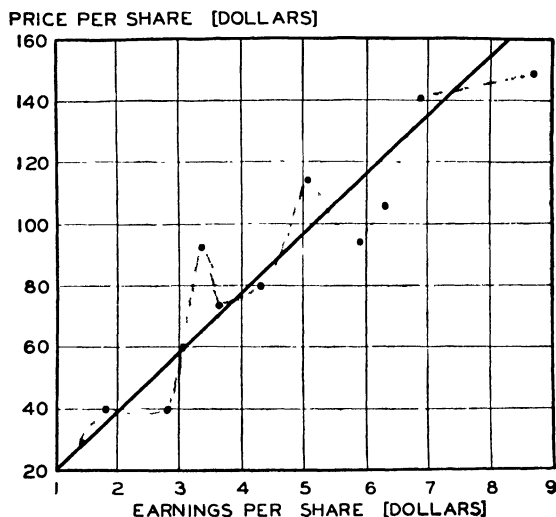
If there is a definite tendency for large values of the X-variable to be paired with large values of the Y-variable, and a similar association appears between small values of the two, the plotted points representing pairs of values of the two variables will resemble Figure 103-A. The scatter of these points falls within a relatively narrow band running from lower left to upper right of the diagram. In most practical examples the number of points is not great enough to produce a clear outline of this elliptical pattern, but a tendency of the points to follow a path upward toward the right denotes a positive association.

When small values of Y are associated with large values of X and vice versa, the pattern will appear in the form shown in Figure 103-B. If either large or small values of Y are found with either large or small values of X, the points will be arranged in the form shown in Figure 103-C. These three diagrams illustrate respectively positive correlation, negative correlation, and no correlation.

Figure 104 shows the nature of the correlation between the earnings per share of common stock of 12 corporations manufacturing industrial chemicals, and the average price of the common stocks of the same corporations. Earnings are plotted on the horizontal scale and stock prices on the vertical. There is a general cause-and-effect relation here in the sense that, other factors being equal, when a corporation is

FIGURE 104

FREEHAND REGRESSION LINE SHOWING RELATION BETWEEN PRICES AND EARNINGS PER SHARE OF COMMON STOCK OF TWELVE CHEMICAL MANUFACTURERS



Data from Table 148, Columns (1) and (2).

earning more, investors will pay more for the right to share in those earnings. However, the other factors are seldom disregarded by investors.

These factors include, (1) the financial strength of the several corporations, (2) the nature of the equity represented by the common stock, (3) the market position of the products, (4) the availability of raw materials, (5) the soundness of research and development programs, (6) the quality of management, and other pertinent circumstances. It would appear, therefore, that a high rate of earnings may or may not cause investors to place a high value on the stock. In spite of the interaction of these and other possible factors that affect either the earnings per share or the price of stock, the location of the points in Figure 104 indicates that in the case of these 12 chemical manufacturers a direct correlation does exist between prices and earnings of common stock.

THE REGRESSION LINE

The relation between the two variables can be indicated by drawing a line on the scattergram.

Free Hand

In the simplest usage a straight line is drawn free hand through the intersection of the two averages, following the path of the points, and located so that the number of points falling on either side will be about equal. To establish the position of the line by inspection it is desirable to place a card, or transparent ruler, in trial positions until that one is found which seems to represent the line of scatter most closely.

This line gives the trend of the relation between the two variables and is known as a "regression line." The designation has survived from the earliest work on correlation by Galton, although the original notion of the regression of hereditary characteristics toward racial norms is not applicable directly to the problems in which regression lines are employed today. Nevertheless the word has survived and we shall continue its use.

By means of the slope of the regression line the average amount of change in one variable can be estimated in terms of the average amount of change in the other. The numerical expression of this relation can be explained by reference to Figure 104. A major part of the X-axis is taken as a base for measuring the slope, and the Y values on the regression line corresponding to the end values of the base distance are read from the vertical scale. Thus the base values of 1 and 7 have corresponding Y values on the regression line of 20 and 135, respectively. The slope of the line is,

$$\frac{135 - 20}{7 - 1} = \frac{115}{6} = 19.17$$

That is, according to this free-hand line an increase of \$19.17 in the price of stocks accompanies an increase of \$1.00 in earnings per share.

Fitted by the Least-Squares Method

A more exact method of locating the regression line is usually employed. It consists in following the same principle that was introduced in chapter XXII, when a free-hand straight-line trend was replaced by a computed trend located by the method of least squares. In correlation analysis no change is required in the least squares method, except to note that the values of X now represent a quantitative attribute classification instead of periods of time. Hence, for ungrouped data, the X values will not be located at regular intervals

along the horizontal scale as was the case when they represented equally spaced intervals of time.

Three Forms of the Equation.—*Form A:* The equation of a straight line fitted to a set of points is,

$$Y = a + bX \quad (1)$$

Specific values of a and b , for a given set of data are obtained by solving the two normal equations (see page 573, chapter XXII):

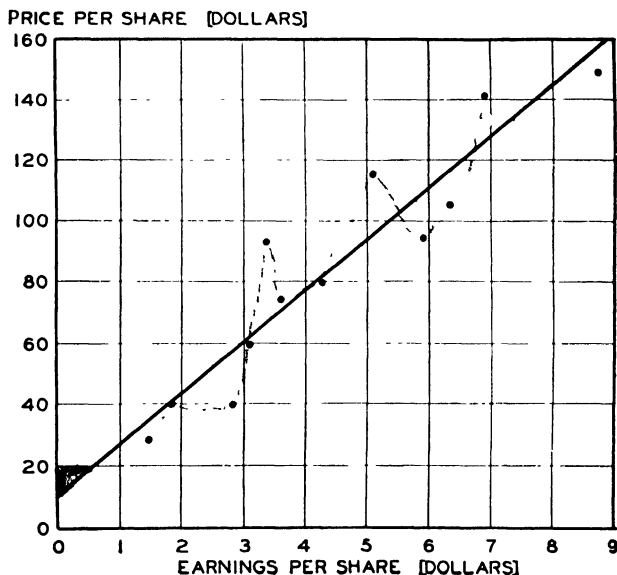
$$\begin{aligned} \Sigma Y &= Na + b \Sigma X \\ \Sigma(XY) &= a \Sigma X + b \Sigma X^2 \end{aligned}$$

FIGURE 105

REGRESSION LINE FITTED BY LEAST SQUARES METHOD TO PRICES AND EARNINGS PER SHARE OF COMMON STOCK OF TWELVE CHEMICAL MANUFACTURERS

A

X and Y measured from 0 origin



Data from Table 148-A.

The computation of the values needed in these equations in fitting a line to earnings and stock prices of chemical manufacturers is given in Table 148-A. Figure 105-A shows the 12 points and the computed regression line. The value of b in the equation is the slope of the line, or the change in price that accompanies one unit of change in earnings.

TABLE 148

THREE FORMS FOR COMPUTATION OF STRAIGHT-LINE TREND (REGRESSION LINE) SHOWING THE RELATION IN 1935 OF AVERAGE PRICES OF COMMON STOCK ON THE NEW YORK STOCK EXCHANGE TO EARNINGS PER SHARE OF TWELVE CHEMICAL MANUFACTURERS

CORPORATION	A USING X AND Y			B USING $\bar{x}$ AND $\bar{y}$ ($\bar{y} = 4.43$)			C USING $\bar{x}$ AND $\bar{y}$ ($\bar{M}_x = 4.43$; $\bar{M}_y = 85.25$)		
	(1) Earnings Per Share $\bar{X}$	(2) Price of Stock $\bar{Y}$	USING X AND Y		(6) $\bar{x}$	(7) $\bar{x}\bar{y}$	(9) $\bar{y}$	(10) $\bar{x}\bar{y}$	(11) $\bar{y}^2$
			(3) XY	(5) Y^2					
Air Reduction	\$ 6.29	\$ 106	666.7	11,236	+ 1.86	+ 197.2	+ 20.75	+ 38.60	430.6
Allied Chemical and Dye	8.71	149	1297.8	22,201	+ 4.28	+ 637.7	+ 63.75	+ 272.85	4064.1
Atlas Powder	2.81	40	112.4	1,600	+ 1.62	+ 64.8	+ 45.25	+ 73.30	2047.6
Dow Chemical	3.35	93	311.6	8,649	+ 1.08	+ 100.4	+ 7.75	+ 8.37	60.1
E. I. Dupont	5.04	116	584.6	13,456	+ .61	+ 70.8	+ 30.75	+ 18.76	945.6
Eastman Kodak	6.90	141	972.9	19,881	+ 2.47	+ 348.3	+ 55.75	+ 137.70	3108.1
Hercules Powder	4.23	80	338.4	6,400	+ .20	+ 16.0	+ 5.25	+ 1.05	27.6
Mathieson Alkali	1.44	29	41.8	841	+ 2.99	+ 86.7	+ 56.25	+ 168.19	3164.1
Monsanto Chemical	3.59	75	269.2	5,625	+ .84	+ 63.0	+ 10.25	+ 8.61	105.1
Parke-Davis	1.82	40	72.8	1,600	+ 2.61	+ 104.4	+ 45.25	+ 118.10	2047.6
Penn. Salt	5.94	94	558.4	8,836	+ 1.51	+ 141.9	+ 8.75	+ 13.21	76.6
Union Carbide and Carbon	3.03	60	181.8	3,600	+ 1.40	+ 84.0	+ 25.25	+ 35.55	637.6
	\$53.15	\$1023	5408.4	103,925	+10.73	+1395.9	+187.50	+885.72	16714.7
					-10.73	-519.3	-187.50	-837	
					0	+ 876.6	0	+ 877.35	
					$1023 = 12a + 0 \times b$ $+ 876.6 = 0 \times a + 52.779b$				
					$a = \frac{1023}{12} = 85.25$ $b = \frac{876.6}{52.779} = 16.61$				
					Regression equation $\bar{Y} = 85.25 + 16.6x$				
					$1023 = 12a + 53.15b$ $5408.4 = 53.15a + 288.17b$ multiplying the first equation by 4.429 and eliminating a by sub- traction $5408.4 = 53.15a + 288.17b$ $4530.9 = 53.15a + 235.40b$ $877.5 = 52.77b$ $b = 16.63$				
					substituting the value of b in either equation, $a = 11.59$				
					Regression equation $\bar{Y} = 11.59 + 16.63x$ or to three significant figures $\bar{Y} = 11.6 + 16.6x$				
					$0 = 12a + 0 \times b$ $877.35 = 0 \times a + 52.779b$ $a = 0$ $b = \frac{877.35}{52.779} = 16.62$				
					Regression equation $\bar{y} = 16.6x$				

The unit of earnings is \$1.00; hence on the average \$1.00 of increase or decrease in the earnings of chemical manufacturers is likely to produce a corresponding change of \$16.60 in stock prices. When $X = 0$, $Y = 11.59$, that is, a gives the value at which the line crosses the Y -axis and is usually referred to as the Y -intercept of the line. In this case the value of a indicates that stocks would sell at about \$11.60 when earnings were zero.

This relation is of doubtful significance, but it serves as a basis for a more general question concerning the interpretation of a regression line beyond the limits of the data to which it is fitted. The problem is akin to that of projecting into future years a trend line fitted to a time series. In chapter XXII we left the way open for the projection of a trend line under certain circumstances. The principal criterion in that case was reasonableness, and the same criterion can be applied to the extension of a regression line. If the nature of the relation between the two variables seems to warrant the assumption that the regression found within given limits will hold beyond those limits, the line may be projected. Unless this assumption can be made, any extension of the regression line should be avoided.

Form B: In fitting trend lines to time series in chapter XXII the X -origin was shifted to the middle time period of the series. The great advantage of this translation of the position of the Y -axis arose from the symmetrical deviations of the time periods about the middle period. The same procedure can be introduced in fitting a regression line, but the advantage no longer applies when the X -variable is an attribute, because its values are usually not evenly spaced. Nevertheless it will be desirable to shift the Y -axis to the arithmetic average of X -values, as an intermediate step toward the subsequent development.

As in preceding chapters, capital letters are used to denote the variables expressed in original form, as in columns 1 and 2 of Table 148, and small letters to denote deviations from an average value as in columns 6 and 9. When deviations are taken from an arbitrary origin the notation d will be used with the subscript x or y .

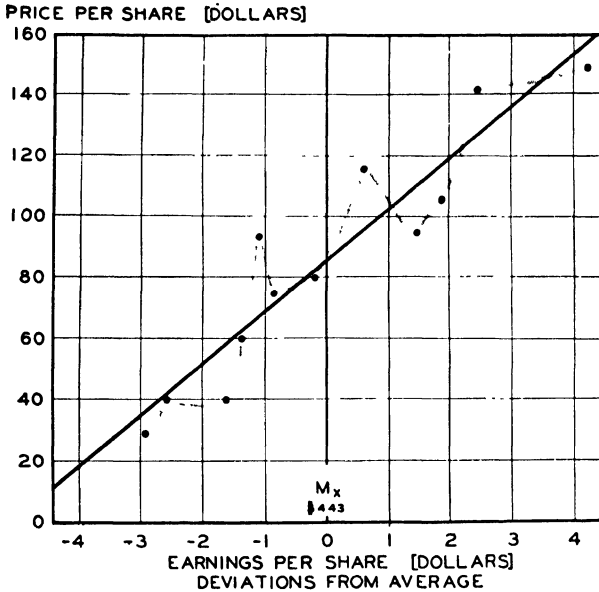
The position of the Y -axis when the origin of x is placed at the X -average is shown in Figure 105-B. The base scale is now written in terms of x , the deviations from the average.¹ The computation of the equation of the regression line is shown in Table 148-B. It should be

¹ In mathematical terms the Y -axis is translated by the equation $X = x + 4.43$.

FIGURE 105 (*continued*)

B

x and Y measured from M_x and 0 origin



Data from Table 148-B.

noted that the solution of the normal equations is greatly simplified by the fact that $\sum x = 0$. The equation is now in the form,

$$Y = a + bx \quad (2)$$

and by substituting values from the table, becomes

$$Y = 85.25 + 16.6x$$

The value of b is the same as in the previous equation (Table 148-A), because the regression lines must be identical by either method when computed from the same data. The value of a is different, since the Y -axis has been moved and the intercept is changed accordingly. The value, $a = 85.25$, indicates that when earnings per share of these chemical manufacturers stand at the average, \$4.43, the price of their stocks will be about \$85.25.

Form C: A second translation of axes consists in moving the X -axis up to the average of Y as shown in Figure 105-C.² The computation of the equation of regression is shown in Table 148-C. The

² The equation of translation is $Y = y + 85.25$.

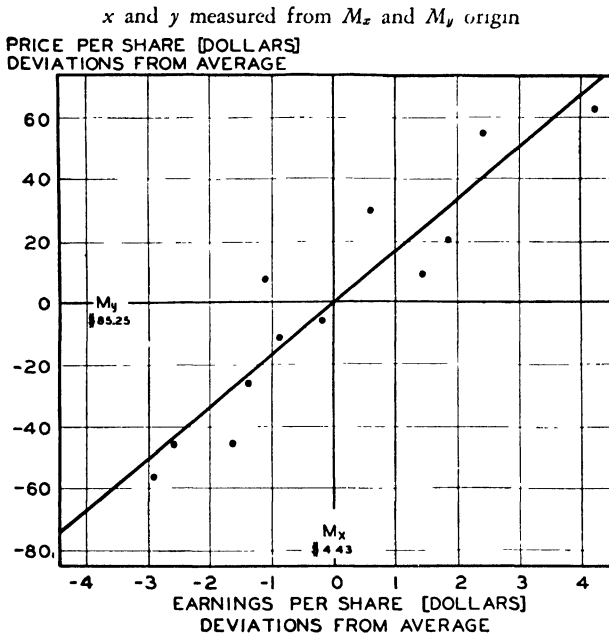
value of a becomes 0 since $\sum y$ is 0 and the equation takes the form,

$$y = bx \quad (3)$$

The value of b continues to be 16.6 because we are dealing with the same regression line that was fitted in Table 148, A and B.

FIGURE 105 (continued)

C



Data from Table 148-C.

Relation between the Three Forms.—The three forms of the regression equation are:

Form A: $Y = 11.6 + 16.6 X$

Form B: $Y = 85.25 + 16.6 x$

Form C: $y = 16.6 x$

The first equation expresses the direct relation between earnings and stock prices with the origin at zero earnings and zero prices. The second equation expresses a mixed relation with the origin at earnings of \$4.43 and zero prices. The third equation expresses the relation between deviations when the origin is at earnings of \$4.43 and prices of \$85.25.

The first form is most convenient for answering questions such as,

What level of stock prices can be expected when earnings are \$5.00? The second form is convenient for the calculation of straight-line trend in a time series, but not for correlation analysis. The third form is used in answering questions such as, What deviation from average stock prices can be expected to accompany a given deviation from average earnings?

The third form of the formula is the simplest to compute and the other forms can be obtained from that one by substitution in the equation. Therefore the work in Parts A and B of Table 148 can be omitted in practice. The third equation can be translated to the form of the first one by the substitution

$$X = x + M_x \quad \text{where} \quad M_x = \frac{\sum X}{N}$$

$$Y = y + M_y \quad \text{where} \quad M_y = \frac{\sum Y}{N}$$

Thus substituting $x = X - 4.429$ and $y = Y - 85.25$, in the equation

$$y = 16.6x,$$

gives $Y - 85.25 = 16.62(X - 4.429)$ ³

Simplifying, $Y - 85.25 = 16.62X - 73.61$

and $Y = 11.6 + 16.6X$

The Uses of Regression

There are two major uses of the regression line. The first is to establish the nature of the relation between the two variables. Whether this relation is direct or inverse is shown by the upward or downward direction of the line. Secondly, the slope of the regression line expresses the average relation between the values of the y and x variables. The value of one variable corresponding to any value of the other may be obtained by solving in the regression equation. Approximately the same operation may be carried out graphically. As previously explained, the value b in the equation expresses this average relation between y and x .

THE STANDARD ERROR OF ESTIMATE

The expression "on the average" has been used repeatedly in connection with the regression line. This usage is based on the resemblance

³ Note that in this substitution it is necessary to carry all computations to four significant figures in order to insure three-significant-figure accuracy in the result.

of the regression line to an average. The average relation between pairs of values corresponds to the average relation indicated by a trend line in a time series. In the case of other averages, representativeness has been determined by the amount of dispersion present. Likewise the validity of the average relation described by a regression line depends upon the amount of dispersion of the points from the line.

The regression line is the best estimate, in the least-squares sense, of the relation of the values of one variable to the values of the other. If by chance all of the plotted points should fall on the regression line, the estimate would be perfect. In general the points will *not* fall on the line, consequently the goodness of estimate is judged in terms of the standard deviation of the points from the regression line. This standard deviation is known as the "standard error of estimate" and is denoted by S_y with a subscript to indicate the variable whose deviations are being measured. It differs from an ordinary standard deviation of a single variable only in that deviations are measured from the regression line instead of from the arithmetic average.⁴

Form of Measuring

At this point in the discussion we are interested in S_y . In symbols the computation is,

$$S_y = \sqrt{\frac{\Sigma(y - y_c)^2}{N}} \quad (4)$$

in which y = ordinates of the plotted points when the origin is at (M_x, M_y)

y_c = ordinates of corresponding points on the regression line, obtained by substituting the given values of x in the equation $y = bx$.

$(y - y_c)$ = the vertical deviations of the plotted points from the regression line.

N = the number of pairs of values of the variables.

Considerable time can be saved by introducing a transformation of the formula. The values of y_c are obtained from the left side of the equation, $y = bx$. Instead of y_c , therefore, bx can be substituted in the formula. The equation becomes,

$$S_y^2 = \frac{\Sigma(y - bx)^2}{N} = \frac{\Sigma(y^2 - 2bxy + b^2x^2)}{N}$$

⁴ The term "standard error" is used in chapter XXIX to describe the dispersion of a particular type of distribution arising in the theory of sampling. To avoid confusion later the student should understand that S_y is a measure of squared deviation from the regression line and *not* a measure of sampling variability.

Writing the summation sign separately with each term gives,

$$S_y^2 = \frac{\Sigma y^2 - 2b \Sigma(xy) + b^2 \Sigma x^2}{N}$$

From the normal equations,

$$b = \frac{\Sigma(xy)}{\Sigma x^2} \quad \text{or} \quad \Sigma x^2 = \frac{\Sigma(xy)}{b}$$

Substituting this value of Σx^2 in the third term on the right we have,

$$\begin{aligned} S_y^2 &= \frac{\Sigma y^2 - 2b \Sigma(xy) + b^2 \frac{\Sigma(xy)}{b}}{N} \\ &= \frac{\Sigma y^2 - 2b \Sigma(xy) + b \Sigma(xy)}{N} \\ &= \frac{\Sigma y^2 - b \Sigma(xy)}{N} \\ S_y &= \sqrt{\frac{\Sigma y^2 - b \Sigma(xy)}{N}}^5 \end{aligned} \quad (5)$$

In this form the value of S_y for earnings and stock prices of chemical manufacturers can be obtained from totals already available in Table 148. Thus,

$\Sigma y^2 = 16714.7$	and
$\Sigma(xy) = 877.35$	$S_y = \sqrt{\frac{16714.7 - (16.62 \times 877.35)}{12}}$
$b = 16.62$	$= \sqrt{177.76}$
$N = 12$	$= 13.3$

Meaning

The standard error of estimate is a measure of the scatter of the points from the regression line. The closer the points lie to the line, the smaller will be the value of S_y and vice versa. S_y is to be interpreted

⁵ This formula may also be used in terms of X and Y . Its form is,

$$S_y^2 = \frac{\Sigma Y^2 - a \Sigma Y - b \Sigma(XY)}{N} \quad (6)$$

Substituting values from Table 148-A gives

$$S_y^2 = \frac{103925 - (11\ 59 \times 1023) - (16\ 63 \times 5408.4)}{12} = 177.23$$

and

$$S_y = 13.3$$

in the same way as any other standard deviation. It gives the range on either side of the regression line within which about 68 per cent of the points can be expected to fall, provided the distributions of both variables are approximately normal. A range of $\pm 2S_y$ should include about 95 per cent of the points and a range of $\pm 3S_y$ should include practically all of the points.

The regression line is a measure of that part of the variability of the y -variable which can be explained by the association of y with x . S_y is a measure of the remaining part of the variability of y , that is *not* explained by the regression line. Obviously if S_y is large, the reliability of the regression line is questionable.

Some standard is necessary, however, in order to determine what constitutes a large or a small value of S_y . The largest value that S_y can take is σ_y . That is, if x and y were completely independent, the regression line of y on x would coincide with the x -axis, and all of the values of y_c would be zero. The computation of S_y would then be identical with the computation of σ_y . Therefore, σ_y is used as the standard for judging whether values of S_y are large or small. In the example previously employed, $\sigma_y = \sqrt{\frac{16714.7}{12}} = 37.3$, S_y was 13.3; hence the scatter about the regression line is roughly one-third as great as the scatter about the average.

This is fairly clear evidence that the regression line is a valuable estimate of the relation between stock prices and earnings per share of chemical manufacturers. The nature of S_y and the relation of S_y to σ_y are shown in Figure 106. The narrow band formed by $\pm S_y$ confirms the previous statement and increases the confidence that can be placed in the regression line fitted to these data.

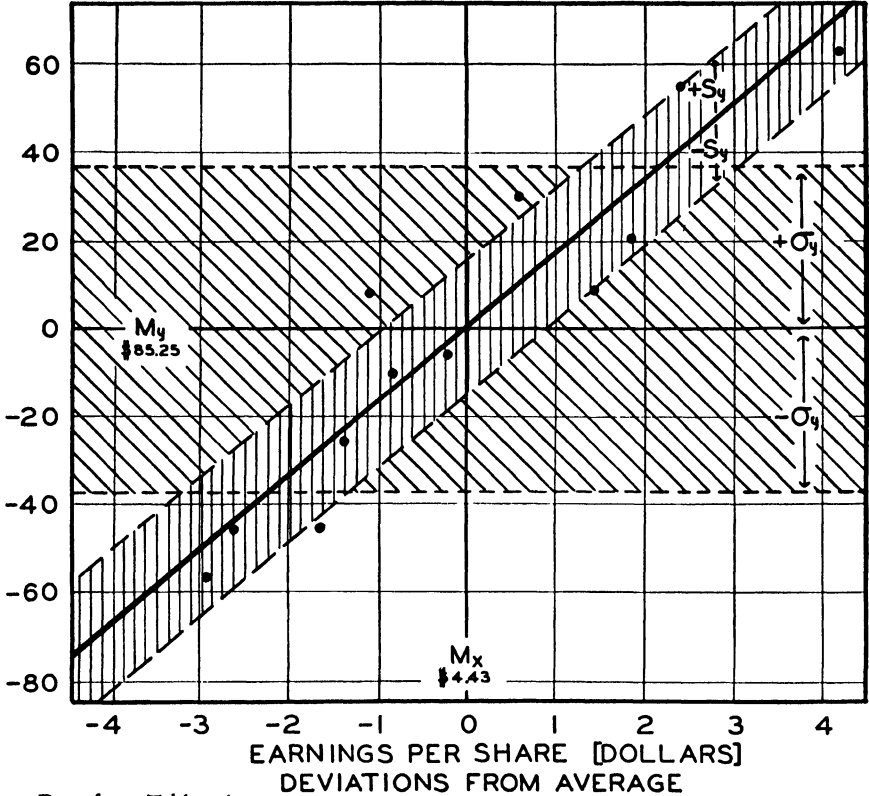
THE COEFFICIENT OF CORRELATION

The ratio of S_y to σ_y provides a numerical measure of the relation between the two variables, y and x . If the two variables are unrelated, S_y and σ_y are identical and the value of $S_y \div \sigma_y$ is unity. If the two variables are perfectly related, i.e., all points falling on the regression line, the value of S_y is zero and $S_y \div \sigma_y$ is zero. As a measure of relation, therefore, the ratio $S_y \div \sigma_y$ possesses the major requirement of having a definite upper and lower limit. But it is inverse in character, its value being large for a low degree of correlation and small for a

FIGURE 106

STANDARD ERROR OF ESTIMATE AND STANDARD DEVIATION OF y , PRICES IN RELATION TO EARNINGS PER SHARE OF COMMON STOCK OF TWELVE CHEMICAL MANUFACTURERS

PRICE PER SHARE [DOLLARS]
DEVIATIONS FROM AVERAGE



Data from Table 148.

high degree of correlation. To avoid this inverse direction of movement and to correspond with subsequent work, the coefficient of correlation (r) is defined by the equation,

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2} \quad (7)$$

When S_y is zero, the value of r is ± 1 . When $S_y = \sigma_y$, the value of r is zero. These limits therefore represent perfect correlation (direct or inverse), and complete absence of correlation. However, this form of the equation does not measure inverse correlation algebraically because S_y and σ_y are always positive. A negative sign must be prefixed to r

when the algebraic sign of b is negative in the regression equation, $y = bx$.

If S_y is not needed as a step in the analysis, the value of r can be obtained more conveniently by changing formula 7 to a form that uses values already available from Table 148-C.

$$\begin{aligned}
 r^2 &= 1 - \frac{S_y^2}{\sigma_y^2} \\
 &= 1 - \frac{\frac{\Sigma y^2 - b \Sigma(xy)}{N}}{\frac{\Sigma y^2}{N}} = 1 - \frac{\Sigma y^2 - b \Sigma(xy)}{\Sigma y^2} \\
 &= \frac{\Sigma y^2 - \Sigma y^2 + b \Sigma(xy)}{\Sigma y^2} \\
 r^2 &= \frac{b \Sigma(xy)}{\Sigma y^2} \quad (8)
 \end{aligned}$$

If the regression line has a negative slope, the sign of b must be prefixed to r as a final step.⁶

The computation of r for stock prices and earnings of chemical manufacturers by the two formulas is as follows:

FORMULA 7	FORMULA 8
$r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$	$r^2 = \frac{b \Sigma(xy)}{\Sigma y^2}$
$= 1 - \frac{(13.3)^2}{(37.3)^2}$	$= \frac{16.62 \times 877.35}{16714.7}$
$= 1 - \frac{177.76}{1392.85}$	$= \frac{14581.56}{16714.7}$
$= 1 - .1276$	$= .8724$
$= .8724$	$r = +.93$
$r = +.93$	

⁶ In footnote 5, p. 716, an alternative formula was given for S_y in terms of X and Y . The corresponding formula can be used for r .

If $r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$ is written as $r^2 = 1 - \frac{\Sigma Y^2 - a \Sigma Y - b \Sigma(XY)}{\Sigma Y^2 - NM_y^2}$
then by simplifying

$$r^2 = \frac{a \Sigma Y + b \Sigma(XY) - NM_y^2}{\Sigma Y^2 - NM_y^2} \quad (9)$$

It should be noted that in this formula and in formula (6) b must be used without regard to sign.

Recapitulation

Three measures of the nature and extent of the association between two variables have been explained in preceding pages. Each of these provides a distinct type of information, but the three are so closely integrated that joint use of them gives a connected analysis of the relation between the paired values of any two such variables. The *regression line* indicates the presence of either positive or negative association. The *standard error of estimate* provides additional information as to the extent of the dispersion of the points from the regression line. The smaller the dispersion, the greater the reliability than can be ascribed to the regression line as a measure of association. But the difference between large and small values of S_y should be expressed numerically. This is done by referring S_y to σ_y as a standard. The particular form of the comparison is designated as the *coefficient of correlation* (r).

Correlation is a measure of the variability of a set of points from the regression line of one variable on the other in relation to the total variability of the points from the average of the dependent variable. The smaller this ratio becomes the greater the reliability of the regression line. Instead of this inverse relation, a direct statement is obtained by writing r as one minus the ratio. The direct form of the coefficient of correlation provides a limited range between $+1$ and -1 for the ratio of association (coefficient of correlation), in which both extremes represent high correlation, with zero correlation at the center.

The connection between the three measures may be stated in another way. The regression line is a measure of the variability in y which is explained by the association of y with x . S_y is a measure of the variability in y not explained by the association of y with x . Then a small value of S_y shows that the regression line is a reliable measure of association and a large value of S_y shows the reverse. S_y , however, is expressed in the units of the y variable, and whether it should be considered large or small depends upon the range of the data. The coefficient of correlation, on the other hand, is an abstract number indicating whether a given value of S_y is high or low, regardless of the values of the data involved. The coefficient of correlation, therefore, is a relative measure of the reliability of an estimate of association found by fitting a line of regression.

Transfer to the Product-Moment Form

For the work which follows a more convenient form of r is desirable. The new form can be obtained by algebraic manipulation of the previous form or by direct logical deduction. The latter will be referred to as the historical development.

Algebraically.—One of the forms previously used in computing the coefficient of correlation was,

$$r^2 = \frac{b \Sigma(xy)}{\Sigma y^2} \quad (\text{formula 8})$$

in which the sign of r will be plus or minus according to the direction of the regression line. As a means of securing a more general expression we now substitute for b its value as obtained from the second normal equation. This equation is

$$\Sigma(xy) = a \Sigma x + b \Sigma x^2$$

But $\Sigma x = 0$, therefore

$$b = \frac{\Sigma(xy)}{\Sigma x^2}$$

Substituting this value of b in formula 8 gives,

$$r^2 = \frac{\Sigma(xy)}{\Sigma x^2} \times \frac{\Sigma(xy)}{\Sigma y^2} = \frac{[\Sigma(xy)]^2}{\Sigma x^2 \times \Sigma y^2}$$

Hence

$$r = \frac{\Sigma(xy)}{\sqrt{\Sigma x^2 \times \Sigma y^2}} \quad (10)$$

But $\Sigma x^2 = N\sigma_x^2$ and $\Sigma y^2 = N\sigma_y^2$
therefore

$$r = \frac{\Sigma(xy)}{\sqrt{N\sigma_x^2 \times N\sigma_y^2}} = \frac{\Sigma(xy)}{N \times \sigma_x \times \sigma_y} \quad (11)$$

This is known as the product-moment form of r . The name is taken from the numerator term which is a sum of products of paired x and y values and can therefore be positive, negative, or zero according to the nature of the association between the two variables. The word "moment" comes from physics and in this case refers to the force exerted about the origin by several pairs of values of (xy) . Thus if the values of a given problem are scattered about the origin in the form of

Figure 105-C, $+x$'s are associated with $+y$'s, and $-x$'s are associated with $-y$'s. Since all such products are plus, the moment or force of the points about the origin will be positive. If in another problem $-y$'s are associated with $+x$'s, and $+y$'s are associated with $-x$'s, the products will be minus and the moment about the origin will be negative. If plus and minus x 's occur with plus and minus y 's at random, $\Sigma(xy)$ will approach zero, the forces in various directions will counterbalance, and the moment will be zero. The sign of $\Sigma(xy)$ determines the sign of r , since the standard deviations in the denominator are necessarily positive.

The product-moment formula will give the same value of r as those previously explained. For example, all of the information needed to compute r for the variables stock prices and earnings per share of chemical manufacturers is available in Table 148. Thus,

$$\Sigma(xy) = +877.35; \quad \sigma_x = \sqrt{\frac{52.779}{12}} = 2.10; \quad \sigma_y = \sqrt{\frac{16714.7}{12}} = 37.3;$$

and

$$r = \frac{+877.35}{12 \times 2.10 \times 37.3} = +.93.$$

Historically.—It will be worthwhile to break the sequential development at this point to introduce a purely logical argument for the use of the product-moment formula as a measure of correlation. This explanation is roughly a description of the discovery and early formulation of the nature of correlation. For this reason it is known as the historical development.

Formulations of the theory of least squares for two variables at the beginning of the nineteenth century brought to light the term, (xy) , and applications of the theory gave the expression, $\Sigma(xy)$. Much later this term was recognized as a measure of association. That is, it was found to give positive, negative, or zero values according to whether two variables were directly or inversely associated or were unrelated. But it was quickly noted that the value of $\Sigma(xy)$ could be compared from one problem to another only if the two had the same number of items, or pairs of values. The next step, therefore, was to shift to the average co-relation per pair of values, and the formula became, $\frac{\Sigma(xy)}{N}$.

This form proved unsatisfactory because the value of the ratio depended upon the amount of dispersion of the variables.

The next step was to express each variable in units of its own standard deviation. Thus the formula became

$$r = \frac{\sum \left(\frac{x}{\sigma_x} \times \frac{y}{\sigma_y} \right)}{N} \quad (12)$$

This form eliminated the several difficulties enumerated and has been used to measure correlation since its discovery sometime prior to 1900. The greater part of the research work implied in this description was either carried out directly by Karl Pearson or was done under his guidance. In recognition of this great contribution the product-moment formula and its variants are always referred to as the "Pearsonian Coefficient of Correlation."

The following manipulation of the preceding expression will produce the exact form of the product-moment formula,

$$r = \frac{\sum \left(\frac{x}{\sigma_x} \times \frac{y}{\sigma_y} \right)}{N} = \frac{\frac{1}{\sigma_x \sigma_y} \sum (xy)}{N} = \frac{\sum (xy)}{N \sigma_x \sigma_y} \quad (\text{formula 11})$$

Correlation in a Cell Table

The frequency distribution was introduced in an earlier chapter as a means of condensing a mass of information so as to make it more convenient for statistical analysis. In the same way a long list of pairs of values of associated variables is usually converted into a double frequency table for convenience in computing the coefficient of correlation. This involves the planning of class intervals for both variables and requires the use of the principles explained in chapter XV concerning number of intervals, width of intervals, and designation of class limits.

Construction of the Table.—In constructing a double frequency, or "cell," table the class intervals of the original X and Y values of the two variables are arranged in the same manner as the scales of the scattergram, Figure 105-A. In Table 149 the earnings per share and stock prices of the 12 chemical manufacturers are used to illustrate on a small scale the method of preparing a cell table. The earnings can be put in \$1.00 class intervals and since there is no tendency toward artificial grouping the limits can be set at the dollar amounts. These classes are shown as the captions of Table 149, corresponding to the X-scale in Figure 105-A. The prices run from \$29 to \$149, and these

have been classified in \$20 intervals in the stub of Table 149. There is no necessity for subdivision into smaller intervals. If there were a large number of items, thirteen \$10 intervals running from \$20 to \$150 would be satisfactory since there is no reason to expect any artificial grouping. The pairs of values are then recorded in the cells by the tallying method as shown in the table. With more items the final step in preparing the table would be to insert in each cell the number of tally marks appearing in that cell. This has been done in the third column as an illustration.

When the tally marks or numbers have been recorded in a cell table, the arrangement and concentration of items is practically the same as in a scattergram. The direction of regression can be located in the cell table by inspection. The X-variable reads from left to right and the

TABLE 149

METHOD OF PREPARING CELL TABLE FROM UNGROUPED DATA—PRICES AND EARNINGS PER SHARE OF 12 CHEMICAL MANUFACTURERS, 1935

PRICE OF COMMON STOCK (Y)	EARNINGS PER SHARE (X)							
	\$1 1 99	\$2-2 99	\$3 3 99	\$4-4 99	\$5-5 99	\$6 6 99	\$7-7 99	\$8-8 99
\$140-\$159 99						I		I
120- 139 99								
100- 119 99					I	I		
80- 99 99			I ①	I	I			
60- 79 99			II ②					
40- 59 99	I	I						
20- 39 99	I							

Y-variable from bottom to top, the same as in the scattergram. If the frequencies tend to concentrate in cells along a path running from lower left to upper right, the correlation is positive. If the frequencies tend to concentrate in cells along a path running from upper left to lower right the correlation is negative. If no tendency toward concentration of frequencies along a diagonal path is discernible, it is likely that the correlation between the two variables is negligible.

When the number of items is small, therefore, it is convenient to use a scattergram and to compute r from the individual items. But when a problem involves a large number of pairs of values of the variables, the construction of a cell table provides a substitute for the scattergram as an indicator of the direction of regression, and at the

same time groups the data in convenient form for the easy computation of r .

Computation of r .—If the arrangement of the frequencies indicates some degree of correlation between the variables, the next step is to compute the value of r . In previous chapters various methods of analysis were illustrated by applying them to the monthly rentals paid by 155 families in Columbus, Ohio. That practice is continued here by computing the coefficient of correlation between these rentals and a second quantitative characteristic of the 155 families, their annual incomes.

The tabulated information appears in the part of Table 150 contained within the double rulings. There is an obvious concentration of frequencies along a diagonal from the lower left corner to the upper right corner of the table. The presence of positive correlation can therefore be assumed, but in order to determine the *extent* of the association, r must be computed.

The grouping of data in class intervals has introduced frequencies, whereas in Table 148-C for ungrouped data the computation was carried out for individual pairs of values of x and y separately. In finding r from a cell table, therefore, the frequencies in the several cells must be provided for in the formula. The product-moment form is the most convenient for this purpose. The $\Sigma(xy)$ becomes $\Sigma(fxy)$, and the standard deviations of x and y in the denominator of r are each computed according to the formula for a frequency distribution as explained in Chapter XVIII.

The formula for r can be written in the form,

$$r = \frac{\Sigma(fxy)}{\Sigma f \times \sqrt{\frac{\Sigma(f_x x^2)}{\Sigma f}} \times \sqrt{\frac{\Sigma(f_y y^2)}{\Sigma f}}} \quad (13)$$

However, the use of M_x and M_y as origins in measuring x and y , respectively will lead to unwieldy figures. To avoid this, assumed averages are used and the notation of previous chapters is adhered to. That is, deviations from the assumed averages are designated as d_x and d_y to distinguish them from x and y , the deviations from the true averages. In symbols the relations are,

M_x = arithmetic average of the X variable

M_y = arithmetic average of the Y variable

TABLE 150

CELL TABLE OR DOUBLE FREQUENCY DISTRIBUTION. COMPUTATION OF COEFFICIENT OF CORRELATION BETWEEN MONTHLY RENTALS AND ANNUAL INCOMES OF 155 FAMILIES IN COLUMBUS, OHIO

MONTHLY RENTS (Y)	ANNUAL INCOME (X)						<i>f_y</i>	<i>d_y</i>	<i>f_yd_y</i>	<i>f_yd_y²</i>	<i>fd_xd_y</i>
	Under \$1,000	\$1,000-2,000	\$2,000-3,000	\$3,000-4,000	\$4,000-5,000	\$5,000-6,000					
\$87.50-97.50						3	3	+5	+ 15	75	+ 45
77.50-87.50				1	6	2	9	+4	+ 36	144	+ 76
67.50-77.50			2	7	1		10	+3	+ 30	90	+ 27
57.50-67.50		1	6	4			11	+2	+ 22	44	+ 6
47.50-57.50		4	10	4			18	+1	+ 18	18	0
37.50-47.50		3	12	2			17	0	0	0	0
27.50-37.50		34	10				44	-1	- 44	44	+ 34
17.50-27.50	5	19	3				27	-2	- 54	108	+ 58
7.50-17.50	16						16	-3	- 48	144	+ 96
<i>f_x</i>	21	61	43	18	7	5	155		+121 -146 - 25	667	+342
<i>d_x</i> ...	- 2	- 1	0	+ 1	+ 2	+ 3					
<i>f_xd_x</i> ...	- 42	-61	0	18	14	15	+ 47 -103 - 56				
<i>f_xd_x²</i> .	84	61	0	18	28	45	236				
<i>f_xd_xd_y</i> .	+116	+66	0	+37	+54	+69	+342				

$c_x = \frac{-56}{155} = -.361$

$c_y = \frac{-25}{155} = -.161$

$$\sigma_x = \sqrt{\frac{236}{155} - (-.361)^2} = 1.18$$

$$\sigma_y = \sqrt{\frac{667}{155} - (-.161)^2} = 2.07$$

$$r = \frac{342 - [(155)(-.361)(-.161)]}{155 \times 1.18 \times 2.07} = \frac{+ 342 - 9.01}{378.6} = \frac{332.99}{378.6} = + .880$$

M'_x = assumed arithmetic average of the X variable
 M'_y = assumed arithmetic average of the Y variable
 $M_x - M'_x = c_x$ = the correction of *x*
 $M_y - M'_y = c_y$ = the correction of *y*
 $x = d_x + c_x$ and $y = d_y + c_y$
 $d_x = x - c_x$ and $d_y = y - c_y$

Then,

$$\begin{aligned}d_x d_y &= (x - c_x)(y - c_y) \\d_x d_y &= xy - c_x y - c_y x + c_x c_y\end{aligned}$$

Multiplying through by f gives,

$$f d_x d_y = fxy - c_x f y - c_y f x + f c_x c_y$$

This is an identical equation expressing for each cell of the table the product of the deviations from the two assumed averages in terms of the deviations from the true averages and the corrections. There will be as many such equations as there are cells in the table and their sum will give the product-moment for the entire table.

$$\Sigma(f d_x d_y) = \Sigma(fxy) - c_x \Sigma(fy) - c_y \Sigma(fx) + c_x c_y \Sigma f$$

but $\Sigma(fy) = 0$ and $\Sigma(fx) = 0$
hence

$$\Sigma(f d_x d_y) = \Sigma(fxy) + c_x c_y \Sigma f$$

Finally

$$\Sigma(fxy) = \Sigma(f d_x d_y) - c_x c_y \Sigma f$$

The two standard deviations are also computed by using the assumed averages according to the formula in chapter XVIII, page 447. The coefficient of correlation becomes

$$r = \frac{\Sigma(f d_x d_y) - c_x c_y \Sigma f}{\Sigma f \times \sqrt{\frac{\Sigma(f d_x^2)}{\Sigma f} - c_x^2} \times \sqrt{\frac{\Sigma(f d_y^2)}{\Sigma f} - c_y^2}} \quad (14)$$

Although this formula has a formidable appearance, the computation of the coefficient of correlation is greatly simplified by its use.

The complete calculation is presented in Table 150. The extension at the right side of the table (except for the last column) is needed in computing σ_y . The assumed average rental is the midpoint, \$42.50, of the fourth class from the bottom, and the d_y column is written in steps. The value of σ_y is therefore expressed in steps, i.e., in units of \$10. The similar extension at the foot of the table for the x -variable leads to the value of σ_x in units of \$1000. The computations below the table are in the usual form for finding c_x , c_y , σ_x , and σ_y .

All of the values needed in computing r are now ready except $\Sigma(f d_x d_y)$. This expression is worked out by one method in the last

column at the right of the table and is recomputed as a check in the bottom row of the table. The computation for the first three rows is,

1st row			2d row			3d row		
f	d_x	d_y	f	d_x	d_y	f	d_x	d_y
$3 \times$	$+3 \times$	$+5 = +45$	$1 \times$	$+1 \times$	$+4 = +4$	$2 \times$	$0 \times$	$+3 = 0$
			$6 \times$	$+2 \times$	$+4 = +48$		$7 \times$	$+1 \times$
			$2 \times$	$+3 \times$	$+4 = +24$		$1 \times$	$+2 \times$
					<u>+76</u>			<u>+27</u>

It will be noted that in each of these three sets the value of d_y is constant. Hence the multiplication, fd_x , can be computed separately for each cell in the row, the products summed, and the result multiplied by d_y . This process can be carried out by mental arithmetic.

The student should practice the mental computation of $fd_x d_y$ in the remaining rows of the table. The computation of $fd_x d_y$ for each column as shown in the last row of the table is a parallel procedure. The value of d_x will be constant for each column. Obviously it is impossible to reproduce on the printed page the exact form of a mental calculation, but the process for the first column might be described somewhat as follows:

$$5 \times -2 = -10; 16 \times -3 = -48; \text{total, } -58 \times -2 = +116$$

The computation of r is completed at the foot of the table by substituting the computed values in formula 14, (page 727). The source of each figure should be evident from the symbols. The entire computation is carried out in step values. There is no reason for introducing the \$1,000 width of the X class interval and the \$10 width of the Y class interval because in so doing we would simply multiply both numerator and denominator by \$10,000. This fact is stressed here in order to point out that when computing the regression line and the standard error of estimate, the values of the standard deviations will not be used in steps but in the original dollar units.

The correlation between income and rent for these 155 families is $+ .88$. This means first of all that the association between the two variables is direct, i.e., as income increases the amount spent for rent also increases. Further, the high value of r indicates that the relation between the two variables holds for most of the cases included. More information on this point can be obtained from the standard error of estimate and the regression equation.

The Standard Error of Estimate.—The original definition of the coefficient of correlation was,

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2} \quad (\text{formula 7})$$

In computing r from a cell table, it was not necessary to compute S_y , therefore by substituting the value of r in this formula, the value of the standard error of estimate, S_y can be obtained.

$$S_y^2 = \sigma_y^2(1 - r^2) \quad (15)$$

Substituting values from Table 150,

$$\begin{aligned} S_y^2 &= (2.07 \times 10)^2[1 - (.88)^2] \\ &= 428.5[1 - .77] \\ &= 428.5 \times .23 \\ &= 98.555 \\ S_y &= 9.93 \end{aligned}$$

The value of S_y shows that about 68 per cent of the rents will fall in a band parallel to the regression line and at a distance of not more than \$9.93 above or below the values on the line, although neither of the variables approaches very closely to a normal distribution.

The Regression Equation.—The information concerning the relation between rent and income is completed by computing the regression equation. The form of the regression equation will be recalled as $y = bx$. The value of b was obtained from the expression $b = \frac{\Sigma(xy)}{\Sigma x^2}$.

In a cell table this becomes $b = \frac{\Sigma(fxy)}{\Sigma(fx^2)}$. Neither of these values is directly available from Table 150 because of the use of assumed averages. It is therefore easier to make an algebraic change in b than to compute these values. From the equation for r (formula 11),

$$\Sigma(fxy) = Nr\sigma_x\sigma_y$$

Substituting this value in b gives

$$b = \frac{Nr\sigma_x\sigma_y}{\Sigma(fx^2)}$$

But

$$\Sigma(fx^2) = N\sigma_x^2$$

hence

$$r = \frac{Nr\sigma_x\sigma_y}{N\sigma_x^2} = r\frac{\sigma_y}{\sigma_x}$$

The regression equation becomes

$$y = r\frac{\sigma_y}{\sigma_x}x \quad (16)$$

but σ_y and σ_x must be expressed in the units of the variables instead of in steps.

For rents in relation to incomes, therefore, the equation is,

$$y = +.88 \frac{2.07 \times 10}{1.18 \times 1000}x$$

or

$$y = +.0154x$$

This equation expresses the relation between deviations from the two averages. For example, if income is \$100 above the average the rent paid will be about \$1.54 above the average, and similarly for any other value substituted for x .

More general use can be made of the equation in terms of X and Y ; therefore we proceed next to shift the origin by the translation,

$$y = Y - M_y; x = X - M_x$$

From Table 150,

$$\begin{aligned} M_y &= 42.50 + (-.161 \times 10) \\ &= 40.89 \end{aligned}$$

$$\begin{aligned} M_x &= 2500 + (-.361 \times 1000) \\ &= 2139 \end{aligned}$$

and

$$\begin{aligned} Y - 40.89 &= +.0154(X - 2139) \\ Y &= 7.95 + .0154X \end{aligned}$$

In this form the equation can be used to determine the rent that families of this group will pay at different income levels. The equation provides an answer to questions such as, What rent would a family pay that has an income of \$1,650? The equation of the regression line is, $Y = 7.95 + .0154X$. Substituting $X = 1,650$, gives $Y = 33.36$. Hence, on the average, rent of \$33.36 accompanies income of \$1,650. Application of the standard error of estimate, \$9.93, to these values shows that in about two-thirds of the cases a family having an income of \$1,650 would be paying rent of not more than \$43.29 and not less than \$23.43.

There are, therefore, two uses of the regression line. First, the rents paid by these 155 families are considered to be a representative sample of rents for the entire city of Columbus, and the regression line, which expresses the average relation in the sample, becomes the basis for inferring the conditions in the universe. This type of inference will receive additional emphasis in chapter XXIX.

The second use is related to the first but has a slightly different purpose. If this sample is to be used to infer conditions in a larger universe, can the association between the two variables be presumed to hold beyond the limits of the values in the sample? For example, is it safe to say that a family with a \$20,000 income will be likely to pay monthly rent of \$316, this being the value found by solving the regression equation? Since \$316 seems to be a rather high rent, it suggests strongly that the regression line is not applicable to incomes as great as \$20,000. There is some doubt as to the use of this line for any value above the \$6,000 upper limit of the sample. This same question was discussed earlier in the chapter (p. 711) in the explanation of the regression line fitted to stock prices and earnings of chemical manufacturers.

SOME DEFERRED POINTS

The preceding discussion has given no hint of any procedures alternative to the finding of three values, (1) the regression line, (2) the standard error of estimate, and (3) the Pearsonian coefficient of correlation. These basic measures should now be sufficiently familiar to permit the mention of several supplementary features of correlation analysis.

The Second Line of Regression and Standard Error of Estimate, S_y

We have spoken of the regression line fitted by making the sums of squares of distances, perpendicular to the x -axis, from the points to the line a minimum. But a second regression line can be fitted by making the sums of squares of distances, perpendicular to the y -axis, from the points to the line a minimum. This line represents the regression of x on y . Its equation is in the same form as that of the first regression line except that the positions of the variables are reversed, i.e.,

$$x = b'y \quad (17)$$

The expression b' is used to indicate that the slope of this line is different from the slope of the first line. The second line is obtained by solving the normal equations with the positions of x and y exchanged.

There is very little reason for using this equation in practical work. If a direct causal or semi-causal relation exists between the two variables, the causal variable would be designated as x and there would be no reason for wishing to know the regression of the cause on the effect. When no causal relation exists between the variables, that one to which the other is to be compared will ordinarily be put on the base line and again no use for the second regression arises. This x on y regression therefore has little use outside of theoretical work. It is introduced merely to fortify the student who may encounter it in further reading or study.

Corresponding to the second regression line there is also a standard error of estimate of the scatter of a given set of points about this line. This is measured by the formula,

$$S_x^2 = \sigma_x^2(1 - r^2) \quad (18)$$

The coefficient of correlation can be obtained from the expression,

$$r^2 = 1 - \frac{S_x^2}{\sigma_x^2} \quad (19)$$

The value of r would be identical whether S_x and σ_x are used or S_y and σ_y .

Other Measures of Correlation

The discussion up to this point has dealt exclusively with Pearsonian r as a measure of correlation. There are, however, many types of association between paired variables in which the regression is non-linear. In all such cases measures other than r must be used. The explanation of such measures lies outside the scope of this book. The correlation index for ungrouped data and the correlation ratio for grouped data are commonly used. Another extensively developed phase of the subject, known as partial and multiple correlation, deals with the simultaneous association of three or more distinct variables.

Several fairly simple measures of correlation have been developed as alternatives to r . Most of these are designed for use with particular types of data. One such, the rank difference method, is explained in the next section.

THE RANK DIFFERENCE MEASURE OF CORRELATION

This method can be used when the paired variable characteristics of a set of items are recorded according to *rank* instead of actual values. That is, for each variable it is more important to know the *position* of each item with respect to the other items than to know the value of each item. When no numerical information is available other than the relative positions of the several items according to two quantitative characteristics, the ranking method of correlation is the only one that can be employed. This situation arises commonly in analyzing data in the field of education. Therefore the rank difference coefficient is a major tool of the educational statistician. However, it is also needed occasionally in analyzing business information.

The basis of the method can be comprehended readily. The items are ranked according to two variable characteristics, either from largest to smallest or the reverse but both must follow the same order. Then if low ranks of one are paired with low ranks of the other, and vice versa, there is direct correlation. If high ranks of one variable are paired with low ranks of the other, the correlation is negative. If there is no tendency to either direct or inverse association of ranks, the two variables are not correlated.

The measure of this correlation denoted by ρ (Greek letter rho), is,

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2 - 1)} \quad (20)$$

in which

D = the difference in rank of paired values
 N = the number of pairs of values

The use of this formula is illustrated in Table 151. The stock prices and earnings of chemical manufacturers that have been used for explanatory purposes earlier in the chapter are also used here. The earnings are ranked from highest to lowest in column 3 and the stock prices are ranked similarly in column 4. The difference in rank of corresponding pairs of the two variables is shown in column 5. Some of the differences are positive and others negative, but the signs can be neglected because the squares in column 6 are all positive.

The third and tenth items in column 4 each have the rank, 10.5. The price of both of these stocks is \$40 and they are the tenth and eleventh in order of rank. There is no way of knowing which stock

should have precedence; therefore each is given the average of the two tied ranks. This is known as the "mid-rank" method of dealing with ties in the rank. If there were three equal values each would be assigned the middle rank of the three ascribable to the items, and similarly for any number of tied values. It should be noted, however, that this is an arbitrary method of resolving ties in the rank. If many ties occur in the values of a given variable, the values should be carried to one more significant figure if possible, in order to eliminate the ties, rather than make extensive use of the mid-rank method.

The computation of the coefficient, ρ , is shown at the bottom of the table. The small value of D^2 is a forecast of a high value of ρ . It turns out to be nearly $+.96$. This can be compared with the value of r , which was $+.93$. The two are distinct methods of measuring correlation; consequently there is no reason why the results should coincide. For data such as those in Table 151 a fairly close correspondence can be expected. Where the values are scattered over a wider range, and particularly when one or the other of the variables includes a few very large values, the difference between ρ and r is likely to be considerably greater than in this example.

TABLE 151

COMPUTATION OF RANK DIFFERENCE COEFFICIENT OF CORRELATION BETWEEN PRICES IN 1935 OF COMMON STOCK AND EARNINGS PER SHARE OF 12 CHEMICAL MANUFACTURERS

(1) EARNINGS PER SHARE X	(2) PRICE OF STOCK ON THE NEW YORK STOCK EXCHANGE Y	(3) RANK OF X	(4) RANK OF Y	(5) DIFFERENCE IN RANK D	(6) DIFFERENCE IN RANK SQUARED D ²
\$6.29	\$106	3	4	1	1
8.71	149	1	1	0	0
2.81	40	10	10.5	.5	.25
3.35	93	8	6	2	4
5.04	116	5	3	2	4
6.90	141	2	2	0	0
4.23	80	6	7	1	1
1.44	29	12	12	0	0
3.59	75	7	8	1	1
1.82	40	11	10.5	.5	.25
5.94	94	4	5	1	1
3.03	60	9	9	0	0
					12.5

$$\rho = 1 - \frac{6 \times 12.5}{12(12^2 - 1)} = 1 - \frac{75}{1716} = 1 - .044 = +.956$$

CORRELATION OF TIME SERIES

Up to this point no attention has been given to the study of the relation between two variables both of which are arranged according to the same time classification. Two questions must be raised in entering upon the use of correlation analysis in the study of time series: (1) What are the limitations upon correlation of time series? (2) What adaptations of previously explained methods of computation are necessary?

Limitations of Series Correlation

In correlating two time series, if the consecutive intervals of time are regarded as a set of items having two quantitative characteristics, the process of correlation can be carried on exactly as for any other pairs of variables. It is not necessary to preserve the order of the years or other time intervals involved so long as each pair of values attaching to the same period is kept together. The variations are not measured from the preceding period but from the average of the entire period, the same as any other x and y deviations. Likewise the variables can be grouped in a cell table. The only additional question involved in time-series correlation is, Which of the several components should become the basis for correlation?

Preceding chapters have shown that the basic analysis of a time series involves the separating of the series into its several components so that the changes in each can be studied individually. Correlation in the study of time series, therefore, implies the relationship between two trends, or two seasonal variations, or two cyclical fluctuations. But other methods already discussed will yield more information about trends and seasonal movements than can be obtained from a coefficient of correlation.

Correlation of time series then means the correlation between the two sets of original data or between the two sets of cyclical fluctuations.

The difficulty in measuring correlation of original data prior to separating them into components is readily apparent from consideration of some of the cases that arise. If one series has a positive trend and the other a negative trend, the correlation between the two series may be negative regardless of how well the cyclical fluctuations agree. If the trends of the two series coincide, a high positive correla-

tion may be found whether or not the cyclical movements agree. (This case is discussed in detail in the succeeding section.) Two series containing little or no trend but containing cyclical movements that agree may have a low correlation because the series contain high-amplitude inverse seasonal movements. These examples are sufficient to indicate that the correlation between two original series may not express the true relation between any parts of the series. Further, there is always a problem of explaining the meaning of such a coefficient.

For cycles, a correlation coefficient is simply a numerical expression of the extent to which the fluctuations of the cycles of the two series tend to occur simultaneously and to be in the same or in opposite directions. A high positive correlation indicates that the two cyclical series tend to move in the same direction. A high negative correlation indicates that the two cyclical series tend to move simultaneously, but that the direction of movement of the cycles is inverse. A low coefficient, either positive or negative, indicates that there is little tendency toward agreement in the cyclical movements of the two series.

Computation of the Coefficient

The computation of correlation of time series by the Pearsonian coefficient requires no change from the form previously used for ungrouped data.⁷ The formula is:

$$r = \frac{\Sigma(xy)}{N\sigma_x\sigma_y} \quad (\text{formula 11})$$

The primary adaptation of the formula consists in defining the two variables x and y differently depending upon which component of the two series is to be correlated. They can be taken to denote the original data or any of the components but most frequently stand for the cyclical fluctuations.

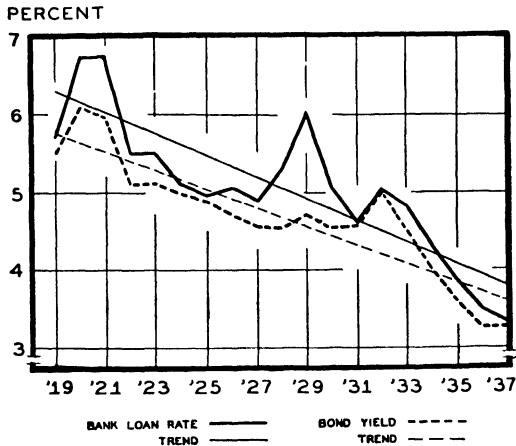
To illustrate the method of computing the coefficient and to provide material for the discussion of several features of series correlation two sets of annual data having nearly parallel trends have been selected. These series are "The Rates Charged by Banks for Customer Loans in Eight Northern and Eastern Cities Exclusive of New York

⁷ In computing the coefficient for series containing as many as 100 pairs of values, some time can be saved by preparing a double frequency or cell table. The calculations would then be identical to those illustrated in Table 150.

City" and "The Yield on Aaa Bonds," for the period 1919–37. The original data are shown in columns 1 and 7 of Table 152-A. The general relation between the two series can be seen in Figure 107.

FIGURE 107

RATES CHARGED BY BANKS FOR CUSTOMER LOANS IN EIGHT NORTHERN AND EASTERN CITIES EXCLUSIVE OF NEW YORK CITY, AND THE YIELD ON AAA BONDS, WITH STRAIGHT LINE TREND FOR EACH SERIES 1919–37



Data from Table 152.

The two appear to have corresponding movements so that a positive correlation is to be expected. In fact they correspond so closely that a question might be raised as to whether the trend and cyclical components need to be treated separately.

Original Data.—In order to facilitate the investigation of this question the computation of the coefficient between the two sets of original data is carried out in Table 152-A. The deviations of the values of each set of data from their respective averages appear in columns 3 and 5. These two columns are labeled x and y since they are in the form for computing the coefficient of correlation. The sum of the products (xy) in column 4 is the numerator of the coefficient. Notice that no corrections are needed since the true average is used in obtaining the deviations in each column.

The two standard deviations are obtained in the usual way from the sum of the squares of deviations in columns 2 and 6. The computation of the denominator of the coefficient of correlation can be abbreviated somewhat by the use of formula 10,

$$r = \frac{\Sigma(xy)}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

TABLE 152

CORRELATION OF ORIGINAL DATA: RATES CHARGED BY BANKS FOR CUSTOMER LOANS IN EIGHT NORTHERN AND EASTERN CITIES
EXCLUSIVE OF NEW YORK CITY AND THE YIELD ON AAA BONDS*

YEARS	A COMPUTATION FOR THE PERIOD 1919-37							B COMPUTATION FOR THE PERIOD 1922-34				
	(1) Bank Loan Rate (per cent)	(2) $\bar{x}^2$	(3) Devia- tions of Loan Rates from Average $\bar{X} = 5.06$	(4) $\bar{xy}$	(5) Devia- tions of Bond Yields from Average $\bar{Y} = 4.68$	(6) $\bar{y}^2$	(7) Bond Yield (per cent)	(8) $\bar{x}^2$	(9) Devia- tions of Loan Rates from Average $\bar{X} = 5.10$	(10) $\bar{xy}$	(11) Devia- tions of Bond Yields from Average $\bar{Y} = 4.72$	(12) $\bar{y}^2$
1919	5.73	.4489	+ .67	+ .5427	+ .81	.6561	5.49					
1920	6.74	2.8224	+1.68	+ 2.4192	+1.44	2.0736	6.12					
1921	6.76	2.8900	+1.70	+ 2.1930	+1.29	1.6641	5.97					
1922	5.48	.1764	+ .42	+ .1764	+ .42	.1764	5.10	.1444	+ .38	+ .1444	+ .38	.1444
1923	5.50	.1936	+ .44	+ .1936	+ .44	.1936	5.12	.1600	+ .40	+ .1600	+ .40	.1600
1924	5.11	.0025	+ .05	+ .0160	+ .32	.1024	5.00	.0001	+ .01	+ .0028	+ .28	.0784
1925	4.98	.0064	— .08	— .0160	+ .20	.0400	4.88	.0144	— .12	— .0192	+ .16	.0256
1926	5.06	0	0	0	+ .05	.0025	4.73	.0016	— .04	— .0004	+ .01	.0001
1927	5.88	.0324	— .18	+ .0198	— .11	.0121	4.57	.0484	— .22	+ .0330	— .15	.0225
1928	4.34	.0784	+ .28	— .0364	— .13	.0169	4.55	.0576	+ .24	— .0408	— .17	.0289
1929	6.04	.9604	+ .98	+ .0490	+ .05	.0025	4.73	.8836	+ .94	+ .0094	+ .01	.0001
1930	5.07	.0001	+ .01	— .0013	— .13	.0169	4.55	.0009	— .03	+ .0051	— .17	.0289
1931	4.61	.2025	— .45	+ .0450	— .10	.0100	4.58	.2401	— .49	+ .0686	— .14	.0196
1932	5.05	.0001	— .01	— .0033	+ .33	.1089	5.01	.0025	— .05	— .0145	+ .29	.0841
1933	4.83	.0529	— .23	+ .0437	— .19	.0361	4.49	.0729	— .27	+ .0621	— .23	.0529
1934	4.29	.5929	— .77	+ .5236	— .68	.4624	4.00	.6561	— .81	+ .5832	— .72	.5184
1935	3.86	1.4400	— 1.20	+ 1.2960	— 1.08	1.1664	3.60					
1936	3.52	2.3716	— 1.54	+ 2.2176	— 1.44	2.0736	3.24					
1937	3.36	2.8900	— 1.70	+ 2.4140	— 1.42	2.0164	3.26					
Totals		15.1615		+12.0926		10.8309	..	2.2826		+ 9937		1.1639

* *Survey of Current Business* (Annual Supplement 1938), pp. 56 and 75.

The value of the coefficient of correlation is obtained by substituting in this formula $\Sigma(xy) = +12.0926$, column 4; $\Sigma x^2 = 15.1615$, column 2; $\Sigma y^2 = 10.8309$, column 6. Thus:

$$r = \frac{+12.0926}{\sqrt{15.1615 \times 10.8309}} = \frac{+12.0926}{12.8146} = +.944$$

The high positive coefficient seems to show that the changes in the two series are to a great extent concurrent. Before this conclusion can be accepted two points must be investigated. The first concerns the components of the two series and the second the specific values appearing in the table. Both of these series have trends with negative slopes; hence the positive deviations of the two series are paired in the early years and the negative deviations are paired in the recent years. The high coefficient therefore is largely a measure of the relation between the two trends. As such it is not particularly valuable and the removal of trend to permit measurement of the correlation of the cycles appears to be in order.

But the values in the table must be studied first. In column 4 the value of $\Sigma(xy)$ is derived almost entirely from the first three and the last three products. Something like this will always occur when a coefficient is computed between two sets of original data with steep parallel trends, but the situation is aggravated here by the presence of positive cycle in the early years and negative cycle in the recent years. The coefficient is completely dominated by these end values and a question naturally arises as to the relation of the two series, exclusive of the end values.

Table 152-B contains the computation of the correlation between the two sets of interest rates for the years 1922-34. The work parallels that in Part A, but the sums of columns 8, 10, and 12 are very much smaller due to the elimination of the six items which formerly contributed most to those sums. The coefficient for the shorter period is:

$$r = \frac{+.9937}{\sqrt{2.2826 \times 1.1639}} = \frac{+.9937}{1.6299} = +.610$$

This result seems to indicate that the movements of the two series are not concomitant to the same extent in the years 1922-34 as for the whole nineteen-year period. But the correlation is still high enough to be interpreted. As in the preceding case, however, the coefficient .61 is a measure of the relation between the combined trends and cycles

of the two series, and almost all of the value of $\Sigma(xy)$ is contributed by the first two and the last two products.

These two examples demonstrate that the correlation of deviations of the original data from their averages is of doubtful significance when applied to time series containing parallel trends. If the trends of the two series had been inverse, a similar situation would have developed except that the coefficients would have been negative. Only in the particular case when neither series contains any appreciable amount of trend will correlations of deviations of original data from the average lead to a coefficient that expresses a significant relation between all of the pairs of values of the two series. In general, then, deviations from trend instead of deviations from the average should be used in computing the correlation between time series.

Cycles.—Direct comparison: The use of deviations from trend to compute the correlation between the two series of interest rates is shown in Table 153-A. Columns 1 and 9 are the original data reproduced from Table 152-A. The straight-line trends fitted to the two series are given in columns 2 and 8. The deviations from trend appear in columns 4 and 6. The products of deviations for the numerator of the coefficient are in column 5, and the squared deviations for the denominator of the coefficient are in columns 3 and 7. Substitution of the sums of these columns in the formula gives the coefficient:

$$r = \frac{+2.4312}{\sqrt{4.4795 \times 2.1165}} = \frac{+2.4312}{3.0791} = +.790$$

This result is specific as contrasted with that obtained by the use of deviations of the original data from their averages. A positive correlation of .79 indicates that the cyclical fluctuations of the two series are predominantly concurrent. Before accepting this conclusion, however, further consideration must be given to the effect of the positive cycles at the beginning of the period and the negative cycles at the end of the period. It will be noted from the table that nearly one-half of the total value of $\Sigma(xy)$ is contributed by the first three and the last three products. The concentration on the end values is not as great in this case as in the two preceding computations, but still presents a somewhat abnormal situation. A survey of the cyclical deviations from trend on the chart shows why the end-year values are so important in $\Sigma(xy)$.

To eliminate the effect of these extreme cyclical fluctuations, the

TABLE 153

CORRELATION OF CYCLES AFTER REMOVAL OF TREND FROM THE DATA OF TABLE 152

YEARS	(1) Bank Loan Rate (per cent) X	(2) Straight Line Trend (per cent) X_t	(3) a^2	(4) Devi- ations of Loan Rates from Trend $X - X_t$ x	(5) xy	(6) Devi- ations of Bond Yields from Trend $Y - Y_t$ y	(7) y^2	(8) Straight Line Trend (per cent) Y_t	(9) Bond Yield (per cent) Y
-------	---	--	--------------	---	-------------	--	--------------	--	--

A
COMPUTATION FOR THE PERIOD 1919-37

1919	5.73	6.32	.3481	-.59	+.1593	-.27	.0729	5.76	5.49
1920	6.74	6.18	.3136	+.56	+.2688	+.48	.2304	5.64	6.12
1921	6.76	6.04	.5184	+.72	+.3240	+.45	.2025	5.52	5.97
1922	5.48	5.90	.1764	-.42	+.1260	-.30	.0900	5.40	5.10
1923	5.50	5.76	.0676	-.26	+.0416	-.16	.0256	5.28	5.12
1924	5.11	5.62	.2601	-.51	+.0816	-.16	.0256	5.16	5.00
1925	4.98	5.48	.2500	-.50	+.0800	-.16	.0256	5.04	4.88
1926	5.06	5.34	.0784	-.28	+.0532	-.19	.0361	4.92	4.73
1927	4.88	5.20	.1024	-.32	+.0736	-.23	.0529	4.80	4.57
1928	5.34	5.06	.0784	+.28	-.0364	-.13	.0169	4.68	4.55
1929	6.04	4.92	1.2544	+.12	+.1904	+.17	.0289	4.56	4.73
1930	5.07	4.78	.0841	+.29	+.0319	+.11	.0121	4.44	4.55
1931	4.61	4.64	.0009	-.03	-.0078	+.26	.0676	4.32	4.58
1932	5.05	4.50	.3025	+.55	+.4455	+.81	.6561	4.20	5.01
1933	4.83	4.36	.2209	+.47	+.1927	+.41	.1681	4.08	4.49
1934	4.29	4.22	.0049	+.07	+.0028	+.04	.0016	3.96	4.00
1935	3.86	4.08	.0484	-.22	+.0528	-.24	.0576	3.84	3.60
1936	3.52	3.94	.1764	-.42	+.2016	-.48	.2304	3.72	3.24
1937	3.36	3.80	.1936	-.44	+.1496	-.34	.1156	3.60	3.26
Totals			4.4795		+2.4312		2.1165	...	...

B
COMPUTATION FOR THE PERIOD 1922-34

1922	5.48	5.46	.0004	+.02	+.0004	+.02	.0004	5.08	5.10
1923	5.50	5.40	.0100	+.10	+.0100	+.10	.0100	5.02	5.12
1924	5.11	5.34	.0529	-.23	-.0092	+.04	.0016	4.96	5.00
1925	4.98	5.28	.0900	-.30	+.0060	-.02	.0004	4.90	4.88
1926	5.06	5.22	.0256	-.16	+.0176	-.11	.0121	4.84	4.73
1927	4.88	5.16	.0784	-.28	+.0588	-.21	.0441	4.78	4.57
1928	5.34	5.10	.0576	+.24	-.0408	-.17	.0289	4.72	4.55
1929	6.04	5.04	1.0000	+.10	+.0700	+.07	.0049	4.66	1.73
1930	5.07	4.98	.0081	+.09	-.0045	-.05	.0025	4.60	4.55
1931	4.61	4.92	.0961	-.31	-.0124	+.04	.0016	4.54	4.58
1932	5.05	4.86	.0361	+.19	+.1007	+.53	.2809	4.48	5.01
1933	4.83	4.80	.0009	+.03	+.0021	+.07	.0049	4.42	4.49
1934	4.29	4.74	.2025	-.45	+.1620	-.36	.1296	4.36	4.00
Totals			1.6586		+3.607		.5219	...	

correlation for the years 1922 to 1934 has been computed in Table 153-B. The several columns contain figures similar to those in corresponding columns of Table 153-A. The coefficient becomes:

$$r = \frac{+.3607}{\sqrt{1.6586 \times .5219}} = \frac{+.3607}{.9304} = +.388$$

The low coefficient brings out for the first time a fact that can be read from Figure 107, namely, the cycles of the two series do not correspond closely except during the end years of the period.

The consecutive steps in correlating the two series of interest rates have been carried out in detail in order to provide a basis for some general observations concerning correlation of time series.

1. Correlation of original data may give misleading results if either or both series contain much trend.

2. Specific interpretation requires that deviations from trend should be used instead of deviations from an average.

3. Unless the amplitudes of all of the cyclical fluctuations of a series are similar, correlation analysis is likely to be misleading.

4. Very careful graphic analysis should precede any attempt to compute a coefficient of correlation.

5. In a large majority of cases graphic analysis will yield as much information as correlation analysis.

6. Percentage deviations instead of amounts of deviation can be used for correlation between cycles. In Tables 153-A and 153-B, columns 4 and 6 could have been expressed as percentages of trend instead of differences from trend. The coefficients obtained by the use of relative deviations would differ slightly from those found by the use of amounts.

7. In dealing with monthly data the trends are fitted to the seasonally corrected figures; otherwise all of the preceding method can be applied without alteration.

Use in measuring lag: The restrictions included in the preceding list are sufficient to indicate that correlation methods have a somewhat limited application in the analysis of time series. The most important use in the past has been to measure the lag existing between two sets of cycles that are related but not simultaneous. When a study of a graph over a light table indicates that the cycles of one series precede the cycles of the other series, the exact number of months of lag present can be determined in some cases by computing several coefficients of correlation.

Suppose that two monthly series, A and B, running from January, 1929, to December, 1939, were being studied, and preliminary analysis indicated that series B lagged several months behind series A. The coefficient of correlation between the cycles would be computed first with no lag, i.e., January, 1929, with January, 1929, February, 1929,

The results of computing several such coefficients, with series B adjusted for one, two, three, four and five months lag, might be as follows:

These coefficients indicate that the cycles of series B occur about three months after the cycles of series A. Further, the relatively high coefficient of $+0.87$ shows that the two sets of cycles correspond very well in both amplitude and period after adjustment has been made for a lag of three periods.

Unfortunately in practice a result as good as this will seldom be encountered. The amplitudes and periods of cycles are too irregular to yield high coefficients and the interpretation becomes difficult on that account. This difficulty is illustrated admirably by an attempt to determine by correlation the exact lag between security prices and interest rates. The cycles of the two series are presented in Figure 97, page 655. As pointed out in the discussion of the chart, interest rates have been so greatly influenced by different economic forces beginning with 1933 that the usual relation between the two curves does not hold in subsequent years. Hence correlation has been used to measure the lag of interest rates during the period 1919 to 1932. Coefficients were computed without lag, with one month of lag adjusted, two months, and so on to eleven months. The detailed work is not reproduced, but the volume of calculation can be surmised from the fact that 168 pairs of values are involved in the first coefficient and 157 pairs in the last one. Several features are apparent from the results shown in Table 154. (1) All of the coefficients are too low to have any exact interpretation, an indication of the lack of correspondence between the two sets of cycles. (2) Despite their

TABLE 154

CORRELATION COEFFICIENTS BETWEEN THE CYCLES OF SECURITY PRICES AND INTEREST RATES, 1919-32, WITH ADJUSTMENT FOR CONSECUTIVE PERIODS OF LAG IN INTEREST RATES

NO. OF MONTHS OF LAG ADJUSTED	COEFFICIENT OF CORRELATION*
None	+ .2656
1 month	+ .2735
2 months	+ .2743
3 months	+ .2856
4 months	+ .2903
5 months	+ .2915
6 months	+ .2905
7 months	+ .2758
8 months	+ .2625
9 months	+ .2416
10 months	+ .2175
11 months	+ .1883

* The coefficients have been carried to four significant figures to show that they form a continuous progression.

low values the coefficients rise gradually to a peak and then decline. If the coefficients were larger, a lag of five months in interest rates would be clearly marked. Even so, the five months' period is strongly suggested by the regularity of the coefficients. (3) The difficulty in using this result can be seen in a discussion of a possible application. Suppose that the cycles of stock prices and interest rates were to be combined in a business index. The cyclical value of stock prices for a given date would be combined with the cyclical value of interest rates for the fifth following month. But the low coefficient means that there is no pronounced tendency for the cycles of stock prices to precede those of interest rates by five months. Therefore it is questionable whether at any given date the use of the five months lag will bring the two sets of cycles into any better conformity than would exist without lagging stock prices. In general, lag based on a coefficient as low as this is scarcely worth using.

In concluding the subject of time-series correlation emphasis must be placed on the exercise of care in the use of a technique that is valuable in some cases but can produce misleading results if applied without discrimination. There is nothing unusual about the examples developed in this section; therefore the difficulties encountered can be taken as typical of what to look for in general. Several rules have been stated, but good judgment is probably a better guide than fixed rules in determining when to proceed with correlation of time series and when to refrain from such action.

RECAPITULATION OF FORMULAS

Regression Lines

Y on X

$$Y = a + bX \quad \text{origin at } (0, 0) \quad (1)$$

$$Y = a + bx \quad \text{origin at } (M_x, 0) \quad (2)$$

$$y = bx \quad \text{origin at } (M_x, M_y) \quad (3)$$

$$y = r \frac{\sigma_y}{\sigma_x} x \quad \text{origin at } (M_x, M_y) \quad (16)$$

$$(y - M_y) = r \frac{\sigma_y}{\sigma_x} (x - M_x) \quad \text{origin at } (0, 0) \quad (\text{see page 730})$$

X on Y

$$x = b'y \quad \text{origin at } (M_x, M_y) \quad (17)$$

Standard Error of Estimate

Y variable

$$S_y = \sqrt{\frac{\Sigma(y - \bar{y})^2}{N}} \quad \text{origin at } (M_x, M_y) \quad (4)$$

$$S_y = \sqrt{\frac{\Sigma y^2 - b \Sigma(xy)}{N}} \quad \text{origin at } (M_x, M_y) \quad (5)$$

$$S_y^2 = \frac{\Sigma Y^2 - a \Sigma Y - b \Sigma(XY)}{N} \quad \text{origin at } (0, 0) \quad (6)$$

$$S_y^2 = \sigma_y^2(1 - r^2) \quad (15)$$

X variable

$$S_x^2 = \sigma_x^2(1 - r^2) \quad (18)$$

Coefficient of Correlation

Pearsonian

For ungrouped data

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2} \quad (7)$$

$$r^2 = 1 - \frac{S_x^2}{\sigma_x^2} \quad (19)$$

$$r^2 = \frac{b \Sigma(xy)}{\Sigma y^2} \quad \text{origin at } (M_x, M_y) \quad (8)$$

$$r^2 = \frac{a \Sigma Y + b \Sigma (XY) - NM_y^2}{\Sigma Y^2 - NM_y^2}$$
 origin at (0, 0) (9)

$$r = \frac{\Sigma(xy)}{\sqrt{\Sigma x^2} \times \sqrt{\Sigma y^2}}$$
 origin at (M_x, M_y) (10)

$$r = \frac{\Sigma(xy)}{N\sigma_x\sigma_y}$$
 origin at (M_x, M_y) (11)

$$r = \frac{\Sigma(\frac{x}{\sigma_x} \times \frac{y}{\sigma_y})}{N}$$
 origin at (M_x, M_y) (12)

In a cell table

$$r = \frac{\Sigma(fxy)}{\Sigma f \times \sqrt{\frac{\Sigma(f_x x^2)}{\Sigma f}} \times \sqrt{\frac{\Sigma(f_y y^2)}{\Sigma f}}}$$
 origin at (M_x, M_y) (13)

$$r = \frac{\Sigma(fd_r d_y) - c_x c_y \Sigma f}{\Sigma f \times \sqrt{\frac{\Sigma(f_x d_x^2)}{\Sigma f} - c_x^2} \times \sqrt{\frac{\Sigma(f_y d_y^2)}{\Sigma f} - c_y^2}}$$
 origin at (M'_x, M'_y) (14)

Rank Difference

$$\rho = 1 - \frac{6 \Sigma D^2}{N(N^2 - 1)}$$
 (20)

PROBLEMS

- 1. In what sense is correlation a more powerful tool of analysis than measures of central tendency and dispersion?
- 2. The sales per employee-hour and salaries and wages as a percentage of sales of 12 variety chain retailers were as follows:

FIRM	SALES PER EMPLOYEE-HOUR (dollars)	SALARIES AND WAGES AS A PERCENTAGE OF SALES
A	1.74	17.7
B	1.74	17.3
C	1.90	16.9
D	2.01	18.8
E	2.05	14.6
F	2.09	18.3
G	2.13	15.0
H	2.17	16.7
I	2.63	16.2
J	2.66	14.1
K	2.76	15.7
L	2.97	14.9

Data adapted from *Expenses and Profits of Limited Price Variety Chains in 1938* by Stanley F. Teele, Harvard University Graduate School of Business Administration, Boston, 1939.

- a) Construct a scattergram of these data.
 - b) Describe the relation between the two variables and explain the reasons for the relation.
 - c) Assuming that salaries and wages per hour were uniform, would all of these points fall on a straight line?
3. a) Compute the equation of regression, the value of S_y and the value of r for the data of Problem 2.
 - b) Discuss in full the meaning of the results obtained in (a).
4. In the text, page 717, the expression "provided the distributions of both variables are approximately normal" is used. What is the reason for inserting this provision?
 5. "The entire subject of correlation could be explained by regression equations."
 - a) Is this statement accurate?
 - b) If so, what is the advantage of introducing r ?
 6. Fifteen different studies of equal importance each furnish (a) the percentage of workers unemployed in one-worker families, and (b) the percentage of workers unemployed in multi-worker families. The fifteen studies are indicated below by the letters A, B, C, etc.

STUDY	PERCENTAGE OF WORKERS UNEMPLOYED IN	
	One-Worker Families	Multi-Worker Families
A	6	11
B	14	27
C	16	28
D	30	45
E	35	50
F	22	33
G	21	30
H	19	27
I	14	21
J	15	21
K	13	21
L	16	24
M	12	18
N	14	21
O	23	32

- a) What correlation is there between unemployment in one-worker families and multi-worker families?
- b) When unemployment is 40 per cent in multi-worker families, what is the probable degree of unemployment in one-worker families?
- c) Is a straight-line regression a permissible assumption in this case?
- d) Can you draw a conclusion as to unemployment in one-worker families when unemployment is 5 per cent in multi-worker families?

7. Is there a cause-and-effect relation between the two variables of Problem 6? Discuss the nature of the relation between these two variables in terms of pages 704-5 of the text.
8. The following is a record of the sales of liquor in a Buffalo store and the temperature, daily, during the months of January, February, and March, 1935.

(NOTE: There is in each week a weekly rhythm in sales which has been eliminated from the figures given below, and which therefore does not enter into your calculations.)

SALES (in dollars)	MEAN TEMPERATURES (in degrees)				
	-10 to 0	0 to +10	+10 to +20	+20 to +30	+30 to +40
100—120 .	3	6	3	...	...
80—100 ...	...	5	12	2	1
60—80 ..	...	3	10	8	1
40-- 60	1	1	6	6	10

- a) Compute the coefficient of correlation between temperature and liquor sales.
 - b) What change in liquor sales can be expected when the temperature varies 5 per cent from the average?
 - c) What amount of sales can this store expect when the temperature is zero? When the temperature is +50 degrees?
 - d) What can you say about extending the regression line beyond the limits of the data?
9. *The Monthly Labor Review* of January, 1941, gives a distribution by age and salary of the 800,000 persons on the civilian federal payroll on December 31, 1938. The following is an approximate percentage distribution of the information:

ANNUAL SALARY (dollars)	AGE (years)				
	20-29	30-39	40-49	50-59	60-69
2,700—2,999	...	1	2	2	...
2,400—2,699	...	3	3	3	1
2,100—2,399	1	4	6	4	2
1,800—2,099	4	6	9	5	2
1,500—1,799	4	4	6	2	1
1,200—1,499	7	5	5	1	...
900—1,199	3	2	2	...	...

- a) Compute the coefficient of correlation between salary and age.
- b) Compute the equation of regression and describe how wage increases are related to age.
- c) Compute the standard error of estimate and state the range of variability of the items from the regression line.

10. Twenty magazine advertisements of nationally known products were shown to 220 persons, 110 men and 110 women. All direct identification of the advertisements was removed so that they could be recognized only by a slogan or a trade mark or a pictorial appeal. For example the Lucky Strike cigarette advertisement was to be recognized by the tobacco leaf.

ADVERTISEMENT	No. OF TIMES RECOGNIZED		ADVERTISEMENT	No. OF TIMES RECOGNIZED	
	By Men	By Women		By Men	By Women
A	29	23	K	79	53
B	84	65	L	35	24
C	91	98	M	64	90
D	38	69	N	31	54
E	76	30	O	40	23
F	101	83	P	76	72
G	82	59	Q	52	39
H	54	70	R	51	64
I	48	30	S	48	63
J	48	15	T	74	27

- a) The figures show that women did not recognize as many advertisements as men, but this may be due entirely to the choice of advertisements. Compute a coefficient of correlation that will answer the question: Do men and women recognize the same advertisements or do some advertisements appeal to men and others to women?
- b) Discuss this question in terms of your computed result.
11. What is the relation of Problem 8 to footnote 7, page 736?
12. Given commodity prices and stock prices for an eighteen-month period.

	COMMODITY PRICES	STOCK PRICES
1936		
1st quarter	80	116
2d quarter	79	121
3d quarter	81	128
4th quarter	83	143
1937		
1st quarter	87	149
2d quarter	88	139

Without lag, $r = +.84$

The general expectation is that stock prices will move about three months in advance of commodity prices. Test the validity of the expectation for the data above. (Trend can be ignored for such a short period. Neither series contains much seasonal variation.)

13. The following table is taken from *Income and Economic Progress*, Volume IV, of The Brookings Institution Studies.

GROSS SALES AND NET PROFITS OF MANUFACTURING CORPORATIONS, 1923-32

YEAR	GROSS SALES (billions of dollars)	PERCENTAGE RATIO OF NET PROFITS TO CAPITALIZATION
1923	54	9.5
1924	51	7.0
1925	57	9.0
1926	60	8.7
1927	61	7.1
1928	64	8.6
1929	69	9.2
1930	58	2.9
1931	42	-1.1
1932	34	-4.9

- a) Compute the coefficient of correlation between sales and the profit ratio.
- b) To what extent has the relationship been obscured because the series have not been separated into their several components?
- c) Compute the equation of regression between the two series.
- d) Interpret (1) the coefficient of correlation; (2) the regression.

REFERENCES

CROXTON, FREDERICK E., and COWDEN, DUDLEY J., *Applied General Statistics*. New York: Prentice-Hall, Inc., 1939.

Chapters XXII-XXV contain a very complete treatment of various phases of correlation.

EZEKIEL, MORDECAI, *Methods of Correlation Analysis*. New York: John Wiley and Sons, 1930.

Recommended for those needing a thorough groundwork in the principles of correlation.

GALE, ARTHUR S., and WATKEYS, CHARLES W., *Elementary Functions and Applications*. New York: Henry Holt and Co., 1920.

The explanation of correlation in pages 420-28 uses a notation which forms the basis for that used in the explanation of a cell table in the text.

KENNEY, JOHN F., *Mathematics of Statistics*. New York: D. Van Nostrand Co., Inc., 1939.

Chapter VIII gives an advanced discussion of correlation theory.

MILLS, FREDERICK C., *Statistical Methods*. New York: Henry Holt and Co., 1938.

Chapters XII, XVI and XVII contain a discussion of correlation measures with non-linear regression and multiple correlation.

SMITH, JAMES G., *Elementary Statistics*. New York: Henry Holt and Co., 1934.

Chapters XX and XXI contain a very lucid explanation of correlation methods. Emphasis is placed on the historical development of the subject.

WALKER, HELEN M., *Studies in the History of Statistical Method*. Baltimore: Williams and Wilkins Co., 1929.

Chapter V contains an excellent resume of the historical development of correlation methods.

YULE, G. UDNY, and KENDALL, M. G., *Introduction to the Theory of Statistics*. London: Charles Griffin and Co., Ltd., 1937.

An elementary explanation of correlation is given in chapter 11 and more advanced parts of the subject appear in chapters 12 and 13. Partial and multiple correlation is discussed in chapters 14 and 15.

CHAPTER XXVIII

THE NORMAL CURVE

THE normal curve occupies such an important place in theoretical and practical statistics that an understanding of at least its simpler properties is an essential part of statistical training. A brief description of the shape of the curve and a graph of it were presented in chapter XV. The measures of central tendency, dispersion, and skewness presented in chapters XVI, XVII, and XVIII were designed to provide additional knowledge of the characteristics of frequency distributions. Figure 64, page 451, in particular shows how the standard deviation and multiples of it divide the total area of the normal curve into segments. The areas of these segments are more fully discussed in this chapter and the next. At this point normal-curve analysis is introduced for two purposes: (1) to amplify further the characteristics of frequency distributions, and (2) to pave the way for the presentation in the next chapter of the principles of sampling.

In chapter XV the normal curve was defined as the shape of the distribution of a large number of measurements around the unknown true value of a physical quantity. That is, the normal curve can be thought of as a description of the occurrence of events affected only by chance. Although the application of the curve in statistics is not primarily to measurements of a physical quantity, the easiest approach to an understanding of the properties of the curve is through their relation to chance. We shall begin, therefore, with a brief statement of the elementary principles of probability.

PROBABILITY

In a single toss of a coin the chance of getting a head is exactly equal to the chance of getting a tail and the probability of each is $1/2$. In rolling a die the probability that any one of the six faces will turn up is $1/6$. In drawing a single card from a deck the probability of the card belonging to any designated one of the four suits is $1/4$. In general, if the happening of a specific thing, i.e., a head on the coin, a six face on the die, or a spade from the deck of cards, is called

success, then the probability of success is the ratio of the number of ways the event (toss of a coin) may *succeed* to the total number of ways the event may happen.

If the first nine digits are written on small discs and the discs are shaken in a vessel, the probability of drawing a five is $1/9$. The probability of drawing less than a five is $4/9$. The probability of drawing not less than a five is $5/9$. A different case arises in computing the probability of obtaining on one draw either the two disc or the eight disc. There are two ways in which the event may succeed and seven ways that it may fail; hence the probability is $2/9$.

The rule is: if an event may succeed in two or more ways the total probability of success is the sum of the two or more separate probabilities. Thus the probability of drawing either a ten spot or a five spot from a deck of cards is $\frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13}$.

Entirely distinct from the probability of "either or" is compound probability, sometimes referred to as "both and." For example, the probability of throwing a four with a die is $1/6$; then the probability of throwing a four on both the first and second throws is the product of the probabilities of the separate events or $1/36$. In the example of the nine discs, the probability of drawing the eight disc and the nine disc in two successive drawings would be $1/81$, if the one first drawn were returned to the vessel before the second drawing, but would be $1/72$ if the first disc were not returned to the vessel. This difference is a direct application of the definition of probability. If the first drawn disc is not returned, only eight discs remain and the probability of drawing a certain one of them is $1/8$.

BINOMIAL DISTRIBUTION

With these general ideas of probability in mind we are ready to deal with the preliminary analysis that leads to an explanation of the normal curve. The first step is a study of the properties of the binomial distribution. If the probability of success in a single event is called p and the probability of failure q , then $(p + q) = 1$. If we are dealing with one event, such as the fall of a single coin, we shall have either one head or one tail; the probability of each being $1/2$.

If two coins are tossed, the fall of a head being considered as a success, the result may be no heads, one head, or two heads, i.e., 0, 1, or 2 successes. The probability of no heads is the probability of two

tails which is $q \times q = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. The probability of one head and one tail is $2 \times (\frac{1}{2} \times \frac{1}{2}) = \frac{1}{2}$ since the event may succeed in either of two ways ($p \times q$, or $q \times p$). The probability of getting two heads is $p \times p = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. A similar argument for the toss of three coins leads to the conclusion that the probabilities of getting no heads, one head, two heads, and three heads are respectively $1/8$, $3/8$, $3/8$ and $1/8$.

These probabilities can be expressed in terms of the binomial expansion and thereby a basis for generalization can be established. If $p = q = \frac{1}{2}$ the relation is

	no heads	PROBABILITY OF one head	two heads	GETTING three heads	four heads	TOTAL NO. OF OF PROB. TOSSED	COINS
$(q + p)^1 = q + p$	$= \frac{1}{2}$	$+ \frac{1}{2}$				$= 1$	1
$(q + p)^2 = q^2 + 2qp + p^2$	$= \frac{1}{4}$	$+ \frac{1}{2}$	$+ \frac{1}{4}$			$= 1$	2
$(q + p)^3 = q^3 + 3q^2p + 3qp^2 + p^3$	$= \frac{1}{8}$	$+ \frac{3}{8}$	$+ \frac{3}{8}$	$+ \frac{1}{8}$		$= 1$	3
$(q + p)^4 = q^4 + 4q^3p + 6q^2p^2 + 4qp^3 + p^4$ etc.	$= \frac{1}{16}$	$+ \frac{4}{16}$	$+ \frac{6}{16}$	$+ \frac{4}{16}$	$+ \frac{1}{16}$	$= 1$	4

The successive terms of the expansion of various powers of $(\frac{1}{2} + \frac{1}{2})$ give the probability of obtaining no heads, one head, etc., in tossing one coin, two coins, etc., the exponent of the binomial being equal in each expansion to the number of coins tossed. For any expansion of the binomial the sum of the probabilities must be equal to unity.

The general formula for the binomial expansion is

$$(q + p)^n = q^n + nq^{n-1}p + \frac{n(n-1)}{1 \times 2} q^{n-2}p^2 + \frac{n(n-1)(n-2)}{1 \times 2 \times 3} q^{n-3}p^3 + \dots + p^n$$

This distribution for n events has the same properties as the specific distributions above.¹

For the immediate development it is desirable to have the expansion of the binomial in integers instead of fractions. To accomplish this a coefficient is added to the binomial and the form becomes

$$2^n(q + p)^n.$$

The expansions then take the form,

	FREQUENCY OF HEADS (SUCCESSES) IN 2 ⁿ TRIALS
	None One Two Three Four Five
$2(q + p)^1 = 2(\frac{1}{2} + \frac{1}{2})^1 = 2(\frac{1}{2} + \frac{1}{2})$	$= 1 + 1$
$4(q + p)^2 = 4(\frac{1}{2} + \frac{1}{2})^2 = 4(\frac{1}{4} + \frac{2}{4} + \frac{1}{4})$	$= 1 + 2 + 1$
$8(q + p)^3 = 8(\frac{1}{2} + \frac{1}{2})^3 = 8(\frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8})$	$= 1 + 3 + 3 + 1$
$16(q + p)^4 = 16(\frac{1}{2} + \frac{1}{2})^4 = 16(\frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16})$	$= 1 + 4 + 6 + 4 + 1$
$32(q + p)^5 = 32(\frac{1}{2} + \frac{1}{2})^5 = 32(\frac{1}{32} + \frac{5}{32} + \frac{10}{32} + \frac{10}{32} + \frac{5}{32} + \frac{1}{32})$	$= 1 + 5 + 10 + 10 + 5 + 1$

¹ This formula will be used here with $p = q = \frac{1}{2}$. Later in the chapter the cases of $p \neq q$ will be discussed.

The terms in the expansion $2^n (\frac{1}{2} + \frac{1}{2})^n$ correspond to the class marks of a frequency distribution and the values of the expansion correspond to the number of frequencies in each class. They can be found quickly by the use of a scheme known as the Pascal triangle. The value of each term in any row is the sum of the term immediately above it and the one to the left of that one. Thus continuing the preceding set,

$$\begin{array}{l}
 32 (\frac{1}{2} + \frac{1}{2})^5 = 1 + 5 + 10 + 10 + 5 + 1 \\
 64 (\frac{1}{2} + \frac{1}{2})^6 = 1 + 6 + 15 + 20 + 15 + 6 + 1 \\
 128 (\frac{1}{2} + \frac{1}{2})^7 = 1 + 7 + 21 + 35 + 35 + 21 + 7 + 1 \\
 256 (\frac{1}{2} + \frac{1}{2})^8 = 1 + 8 + 28 + 56 + 70 + 56 + 28 + 8 + 1 \\
 512 (\frac{1}{2} + \frac{1}{2})^9 = 1 + 9 + 36 + 84 + 126 + 126 + 84 + 36 + 9 + 1 \\
 1024 (\frac{1}{2} + \frac{1}{2})^{10} = 1 + 10 + 45 + 120 + 210 + 252 + 210 + 120 + 45 + 10 + 1 \\
 2048 (\frac{1}{2} + \frac{1}{2})^{11} = 1 + 11 + 55 + 165 + 330 + 462 + 462 + 330 + 165 + 55 + 11 + 1 \\
 4096 (\frac{1}{2} + \frac{1}{2})^{12} = 1 + 12 + 66 + 220 + 495 + 792 + 924 + 792 + 495 + 220 + 66 + 12 + 1 \\
 \text{etc.}
 \end{array}$$

Certain of the features of these expansions will be needed in the subsequent explanation. The number of terms in the expansion is one greater than the power of the binomial. The expansions of even powers have a maximum term at the center. The expansions of odd powers have two equal greatest terms at the center. Equidistant terms on either side of the center term or pair of terms are identical in value.

Computation of the arithmetic average and standard deviation of a few of these distributions will lead to general expressions of key importance. This work is carried out in Table 155. In each distribution the arithmetic average obtained by calculation is equal to np and likewise each standard deviation squared is equal to npq . The general nature of these formulas can be inferred from the examples given. However, the following algebraic proofs will serve to reinforce the logical inference.

The values of the terms of the binomial $(q + p)^n$ are the frequencies of the class intervals, 0, 1, 2, etc. heads, in the coin flipping example. Since these are discrete data, the numbers of heads become the midpoints of the intervals and the average number of heads is found by multiplying each class value by its corresponding frequency, summing and dividing by the total frequency. Since the total frequency in this case is one, the last step is unnecessary. The work can be set up horizontally more easily than in the usual columns.

$$\begin{aligned}
 M = 0 \times q^n + 1 \times nq^{n-1}p + 2 \times \frac{n(n-1)}{1 \times 2} q^{n-2}p^2 \\
 + 3 \times \frac{n(n-1)(n-2)}{1 \times 2 \times 3} q^{n-3}p^3 + \dots + np^n
 \end{aligned}$$

$$= np \left[q^{n-1} + (n-1)q^{n-2}p + \frac{(n-1)(n-2)}{1 \times 2} q^{n-3}p^2 + \dots + p^{n-1} \right]$$

but the expression in the brackets is $(q + p)^{n-1}$

therefore $M = np[q + p]^{n-1} = np$

The standard deviation was shown on page 446 to be $\sqrt{\frac{\Sigma(fX^2)}{\Sigma f} - M^2}$

In symbols,

$$\begin{aligned} \sigma^2 &= \left[0^2 \times q^n + 1^2 \times nq^{n-1}p + 2^2 \times \frac{n(n-1)}{1 \times 2} q^{n-2}p^2 \right. \\ &\quad \left. + 3^2 \times \frac{n(n-1)(n-2)}{1 \times 2 \times 3} q^{n-3}p^3 + \dots + n^2 p^n \right] - (np)^2 \\ &= np \left[q^{n-1} + 2(n-1)q^{n-2}p + 3 \frac{(n-1)(n-2)}{1 \times 2} q^{n-3}p^2 \right. \\ &\quad \left. + 4 \frac{(n-1)(n-2)(n-3)}{1 \times 2 \times 3} q^{n-4}p^3 + \dots + np^{n-1} \right] - (np)^2 \\ &= np \left[\left\{ q^{n-1} + (n-1)q^{n-2}p + \frac{(n-1)(n-2)}{1 \times 2} q^{n-3}p^2 + \dots + p^{n-1} \right\} \right. \\ &\quad \left. + (n-1)p \left\{ q^{n-2} + 2 \frac{(n-2)}{1 \times 2} q^{n-3}p + 3 \frac{(n-2)(n-3)}{1 \times 2 \times 3} q^{n-4}p^2 \right. \right. \\ &\quad \left. \left. + \dots + p^{n-2} \right\} \right] - (np)^2 \\ &= np \left[\{ q + p \}^{n-1} + (n-1)p \left\{ q^{n-2} + (n-2)q^{n-3}p + \frac{(n-2)(n-3)}{1 \times 2} q^{n-4}p^2 \right. \right. \\ &\quad \left. \left. + \dots + p^{n-2} \right\} \right] - (np)^2 \\ &= np[1 + (n-1)p\{q + p\}^{n-2}] - n^2 p^2 \\ &= np[1 + np - p] - n^2 p^2 \\ &= np + n^2 p^2 - np^2 - n^2 p^2 \\ &= np(1 - p) = npq \end{aligned}$$

and

$$\sigma = \sqrt{npq}$$

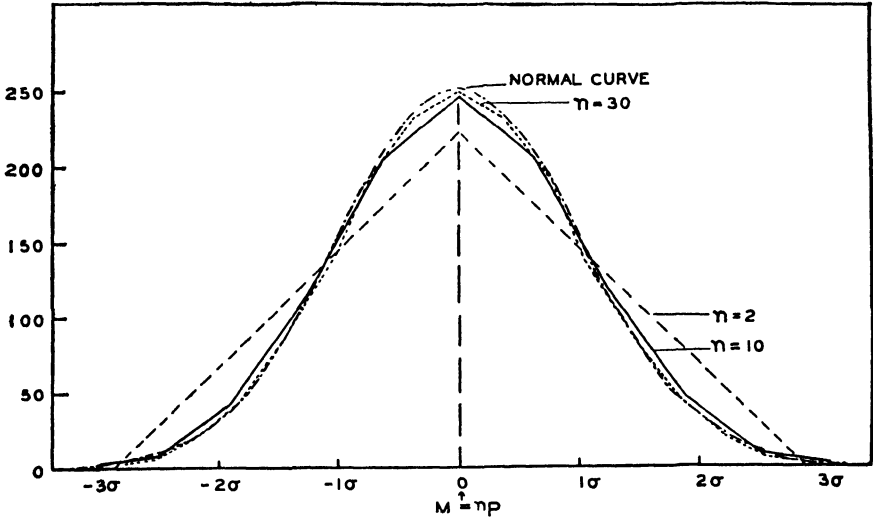
A further generalization of the binomial distribution consists in substituting a general N for 2^n so that the form is written $N(q + p)^n$. The total value of the terms of the expansion will then be equal to N . Since the values of the terms of the expansion are the frequencies of a

distribution with $(n + 1)$ terms or classes, the total frequency will be N . In this form binomial expansions for different values of n can be compared or a binomial expansion can be fitted to a given frequency distribution. These two processes will now be demonstrated.

FIGURE 108

CURVES OF THREE BINOMIAL EXPANSIONS COMPARED WITH NORMAL CURVE

FREQUENCIES



Data for binomial distributions from Table 156.

For Various Values of n

Figure 108 shows the curves of the expansions of

$$\begin{aligned} 1000 (q + p)^2 \\ 1000 (q + p)^{10} \\ 1000 (q + p)^{30} \end{aligned}$$

when $q = p = \frac{1}{2}$.

The curves are plotted with their respective values of np at the center origin and with base scales equated so that the values of $\sqrt{npq}$ coincide for all of the distributions. The steps in this computation are as follows: (1) The class marks of each distribution are expressed in units of the standard deviation of that distribution; (2) the value of the standard deviation of one distribution is selected as a "standard," and the standard deviations of the other distributions are equated to it on the x scale; (3) the ratio of the standard deviation of each distribu-

TABLE 155
COMPUTATION OF THE ARITHMETIC AVERAGE AND STANDARD DEVIATION OF
DISTRIBUTIONS OF NUMBER OF HEADS APPEARING IN COIN TOSSING

NO. OF HEADS	BINOMIAL FREQUENCY	DEVIATION FROM ASSUMED AVERAGE			
	<i>f</i>	<i>d</i>	<i>fd</i>	<i>fd</i> ²	
A Distribution of 4 ($\frac{1}{2} + \frac{1}{2}$) ²					
0	1	- 1	- 1	1	$M' = 1$
1	2	0	0	0	$M = 1 + \frac{0}{4} = 1$
2	1	1	1	1	$\sigma^2 = \frac{2}{1} = \frac{1}{2}$
Total . . .	4		0	2	$npq = 2 \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{2}$ $np = 2 \times \frac{1}{2} = 1$
B Distribution of 128 ($\frac{1}{2} + \frac{1}{2}$) ⁷					
0	1	- 4	- 4	16	$M' = 4$
1	7	- 3	- 21	63	$M = 4 - \frac{64}{128}$
2	21	- 2	- 42	84	$= 4 - .5$
3	35	- 1	- 35	35	$= 3.5$
4	35	0	0	0	$\sigma^2 = \frac{256}{128} - (.5)^2$
5	21	1	21	21	$= 2 - 25$
6	7	2	14	28	$= 1.75$
7	1	3	3	9	$npq = 7 \times \frac{1}{2} \times \frac{1}{2} = 1.75$ $np = 7 \times \frac{1}{2} = 3.5$
Total . . .	128		- 64	256	
C Distribution of 4096 ($\frac{1}{2} + \frac{1}{2}$) ¹²					
0	1	- 6	- 6	36	$M' = 6$
1	12	- 5	- 60	300	$M = 6$
2	66	- 4	- 264	1056	$\sigma^2 = \frac{12288}{4096} = 3$
3	220	- 3	- 660	1980	$npq = 12 \times \frac{1}{2} \times \frac{1}{2} = 3$
4	495	- 2	- 990	1980	$np = 12 \times \frac{1}{2} = 6$
5	792	- 1	- 792	792	
6	924	0	0	0	
7	792	1	792	792	
8	495	2	990	1980	
9	220	3	660	1980	
10	66	4	264	1056	
11	12	5	60	300	
12	1	6	6	36	
Total . . .	4096	.	0	12288	

tion to this "standard" is used as a factor by which the frequencies of the several distributions are multiplied, in order to obtain equivalent areas on the diagram. This process is a direct application of the method demonstrated in chapter XV, Table 67 and Figure 55, by which the frequencies in a *single* distribution having intervals of unequal width are adjusted to represent correctly the area under the curve.

Table 156 contains the computations by which the values are ob-

tained for the three binomial distributions, $1000(q + p)^2$, $1000(q + p)^{10}$, and $1000(q + p)^{30}$, and their curves are plotted in Figure 108 in comparison with a normal curve. The expansion of $1000(q + p)^{10}$ is taken as the standard, and its curve, therefore, is plotted directly from columns 2 and 3 in Part A of the table. The abscissas of $1000(q + p)^2$ are plotted from Part B, column 2, and the adjusted ordinates from column 4. In like manner the curve of $1000(q + p)^{30}$ is plotted from columns 2 and 4 of Part C.

TABLE 156

COMPUTATIONS FOR FIGURE 108, REDUCING THE EXPANSIONS OF THREE BINOMIALS TO A COMPARABLE BASIS

A. DISTRIBUTION OF $1000(\frac{1}{2} + \frac{1}{2})^{10}$

(1)	(2)	(3)
x	$\frac{x}{\sigma}$	f
-5 .. .	-3 16	.98
-4	-2 53	9 77
-3	-1 90	43 95
-2	-1 27	117 19
-1... ..	-.63	205 08
0	0	246 09
+1 ..	+.63	205 08
+2 ..	+1 27	117 19
+3	+1 90	43 95
+4	+2 53	9 77
+5 .	+3 16	.98
		1000 0

$$M = np = 5.$$

$$\sigma = \sqrt{npq} = \sqrt{\frac{10}{4}} = \sqrt{2.5} = 1.58$$

B DISTRIBUTION OF $1000(\frac{1}{2} + \frac{1}{2})^2$

(1)	(2)	(3)	(4)
x	$\frac{x}{\sigma}$	f	$f \times \frac{\sigma^2}{\sigma_{10}} = f \times .4475$
-1.	-1 414	250	111 88
0	0	500	223 75
+1.....	+1 414	250	111 88
		1000	447.51

$$M = np = 1;$$

$$\sigma = \sqrt{npq} = \sqrt{\frac{1}{2}} = .7071$$

C DISTRIBUTION OF $1000(\frac{1}{2} + \frac{1}{2})^{30}$

(1)	(2)	(3)	(4)
x	$\frac{x}{\sigma}$	f	$f \times \frac{\sigma_{30}}{\sigma_{10}} = f \times 1.734$
-15 ..	-5 47	0	
-14 . .	-5 11	0	
-13 .	-4 74	0	
-12 .	-4 38	0	
-11 .	-4 01	.03	.05
-10 .	-3 65	.13	.23
-9 .	-3 28	.55	.95
-8 .	-2 92	1 90	3 29
-7 .	-2 55	5 45	9 45
-6 .	-2 19	13 52	23 10
-5 .	-1 82	27 98	48 52
-4 .	-1 46	50 88	88 23
-3 .	-1 09	80 55	139 67
-2 .	-.73	111 54	193.41
-1 .	-.36	135 44	234 85
0 .	0	144 46	250 49
+1 .	+.36	135 44	234 85
+2 .	+.73	111 54	193 41
+3	+1 09	80 55	139 67
+4 .	+1 46	50 88	88 23
+5 .	+1 82	27 98	48 52
+6	+2 19	13 52	23 10
+7	+2 55	5 45	9.45
+8	+2 92	1 90	3.29
+9 .	+3 28	.55	.95
+10	+3 65	.13	.23
+11 .	+4 01	.03	.05
+12	+4 38	0	
+13	+4 74	0	
+14	+5.11	0	
+15.....	+5 47	0	
		1000 0	1733 99

$$M = np = 15;$$

$$\sigma = \sqrt{npq} = \sqrt{\frac{30}{4}} = 2.74$$

The distribution of $(q + p)^2$ has only three terms; hence the "curve" representing it is a polygon of four sides (excluding the base). The "curve" of $(q + p)^{10}$ is a polygon of twelve sides. It begins to take the characteristic bell-shape. The "curve" of $(q + p)^{30}$ gives even greater evidence of the approach of the binomial to the normal form as n is increased. The closeness of this approach is indicated by comparing $(q + p)^{30}$ with the normal curve included on the chart.

TABLE 157
BINOMIAL DISTRIBUTION FITTED TO A FREQUENCY DISTRIBUTION OF MONTHLY COST OF ELECTRIC CURRENT OF A DOMESTIC CONSUMER FOR A TEN-YEAR PERIOD

MONTHLY COST	NO. OF MONTHS	$120(\frac{1}{2} + \frac{1}{2})^{12}$
\$2.70—2.79	0	.03
2.80—2.89	1	.35
2.90—2.99	2	1.93
3.00—3.09	6	6.45
3.10—3.19	12	14.50
3.20—3.29	22	23.20
3.30—3.39	31	27.07
3.40—3.49	19	23.20
3.50—3.59	14	14.50
3.60—3.69	5	6.45
3.70—3.79	4	1.93
3.80—3.89	2	.35
3.90—3.99	2	.03
	120	119.99

Fitted to a Given Distribution

The fitting of the binomial $N(\frac{1}{2} + \frac{1}{2})^n$ to a given distribution is shown in Table 157. The monthly electric bills of a domestic consumer vary from month to month for a number of reasons. The meter is not always read on the same date. The months vary in length. More lighting current is consumed in winter than in summer. A refrigerator requires more current in summer than in winter. The family is sometimes away from home for a week in summer. All of these factors tend to spread the monthly bills over a wide range. But running through all these dispersing factors is the persistent tendency of a family to use a constant amount of current for such things as the washing machine, toaster, iron, sweeper, and similar equipment. The net effect of all of these influences seems to be a distribution in about the same form as has been observed for the chance variations of a set of physical measurements. This apparent resemblance will be tested by fitting a binomial distribution to the set of electric bills. The binomial is intended for cases in which two mutually exclusive

occurrences of an event are possible. In fitting a binomial to a distribution of costs of electric current the two occurrences would be a monthly bill less than that at the center of the modal class (probability q) and a monthly bill equal to or greater than that at the center of the modal class (probability p).

In preparing the binomial the power of the expansion to use can be determined by taking twice the number of class intervals on the "long" side of the distribution. In this case the twelfth power should be used. The eleventh power would give two equal frequencies at the center whereas the actual distribution appears to have a well-defined modal class at \$3.30 to \$3.40. The twelfth power was chosen so as to include all of the intervals that contained any frequencies.

The terms of the binomial coincide with the midpoints of the classes of the given distribution, the center of the binomial being located at the midpoint of the modal class. The fitted frequencies differ somewhat from the actual ones. The fit is better for classes less than the interval containing the maximum frequency than for those above that interval. The concentration in the center interval is greater in the actual distribution than in the binomial.

While the fitting of the binomial frequencies conveys some idea of the relation of the given distribution to the normal, the comparison is far from definitive because the binomial does not coincide closely with the true normal form until n becomes 30 or greater. The fitting of binomials with as many as 31 terms is laborious. Moreover the binomial frequencies are essentially discrete values and the curve obtained by joining them is not sufficiently flexible for work explained later in the chapter. If a given distribution is slightly asymmetrical, a normal curve fitted to it by means of the tables introduced in the next section will adjust to this condition. But a binomial is always symmetrical on the two sides of the modal class or classes as long as $p = q = .5$. Therefore the use of the binomial with $p = q = .5$ should be confined to distributions that are very close to the symmetrical normal form. For all of these reasons the binomial frequencies are not commonly used in practical work.

THE NORMAL CURVE

Equation and Properties

The study of Figure 108 has indicated that as n is increased in the expression $N(\frac{1}{2} + \frac{1}{2})^n$, the shape of the binomial approaches the nor-

mal curve. It is natural therefore to go to the limiting case to obtain an equation of the curve. This approach produces an equation which is presented here without proof. The equation is used in several forms according to the types of mathematical work contemplated in a particular case, but the form most useful is,

$$Y = Y_0 e^{-\frac{x^2}{2\sigma^2}} \text{ or } Y = \frac{Y_0}{e^{\frac{x^2}{2\sigma^2}}}$$

The notation of the formula conforms to the usage of preceding chapters. The Y 's are measured from a horizontal base such that all values of Y will be positive. The x 's are deviations from the arithmetic average of the values on the base line. Since x appears as an exponent, the equation is not similar to any of those given in previous chapters. However, a simple analysis will show the general shape of the curve. Since x is squared, the value of Y is the same for $-x$ as for $+x$. Therefore the curve is symmetrical about the Y -axis. By substituting values for x , it can be seen that the expression $e^{\frac{x^2}{2\sigma^2}}$ has the minimum value unity when $x=0$. Hence the right-hand side of the equation is greatest when $x=0$, i.e., Y_0 is the maximum ordinate. The constant e is the base of the Naperian or natural system of logarithms and has the value 2.71828...

By methods of the integral calculus it can be shown that

$$Y_0 = \frac{N}{\sigma\sqrt{2\pi}}$$

where N is the frequency of a distribution to which a normal curve is to be fitted, $\sqrt{2\pi} = 2.5066$ and the value of σ is obtained from the given distribution.² The complete equation is,

$$Y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

Method of Fitting

By Use of the Equation.—The general shape of the normal curve can be shown by plotting pairs of values of x and Y obtained in the usual way.³ Let $N=1$ and $\sigma=1$ in the equation; then

² When computing Y_0 for use in fitting a normal distribution to an actual distribution by the Table of Ordinates, σ must be divided by the width of the class interval unless it has been computed in steps.

$$Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$\text{If } x = 0, Y = \frac{1}{\sqrt{2\pi}} = .39894$$

$$x = \pm \frac{1}{4}, Y = \frac{1}{\sqrt{2\pi}} \times \frac{1}{e^{\frac{1}{32}}} = .39894 \times .96924 = .38667$$

$$x = \pm \frac{1}{2}, Y = \frac{1}{\sqrt{2\pi}} \times \frac{1}{e^{\frac{1}{8}}} = .39894 \times .88250 = .35207$$

$$x = \pm 1, Y = \frac{1}{\sqrt{2\pi}} \times \frac{1}{e^{\frac{1}{2}}} = .39894 \times .60653 = .24197$$

$$x = \pm 1\frac{1}{2}, Y = \frac{1}{\sqrt{2\pi}} \times \frac{1}{e^{\frac{9}{8}}} = .39894 \times .32465 = .12952$$

$$x = \pm 2, Y = \frac{1}{\sqrt{2\pi}} \times \frac{1}{e^2} = .39894 \times .13534 = .05399$$

$$x = \pm 2\frac{1}{2}, Y = \frac{1}{\sqrt{2\pi}} \times \frac{1}{e^{\frac{25}{8}}} = .39894 \times .04394 = .01753$$

$$x = \pm 3, Y = \frac{1}{\sqrt{2\pi}} \times \frac{1}{e^{\frac{9}{2}}} = .39894 \times .01111 = .00443$$

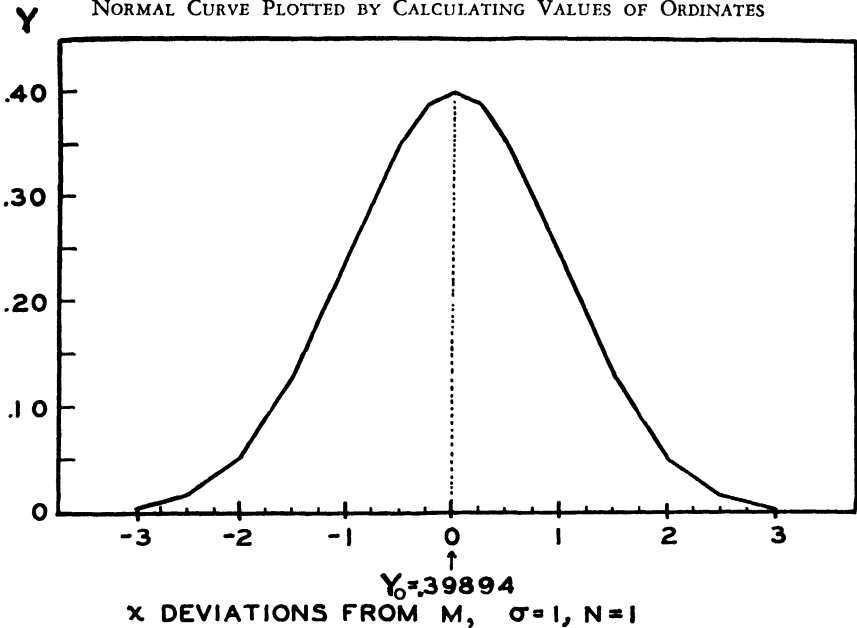
These values of x and Y are plotted in Figure 109. By increasing the number of points the precise curvature of all parts of the curve could be reproduced, but the diagram indicates closely enough the characteristic bell-shape of the normal curve. These values of Y are not frequencies but measures of the *rate* at which frequencies occur at the several values of x . Obviously the sum of the values of Y computed for a large number of values of x would exceed the assumed total frequency of unity.

It is not necessary to derive these values of Y from the equation each time they are wanted. A table of ordinates has been prepared for the purpose and is reproduced in Appendix E.

By Use of Table of Areas.—The process of fitting a normal curve by any method depends upon three assumptions: (1) that the center of the normal distribution shall coincide with the arithmetic average

FIGURE 109

NORMAL CURVE PLOTTED BY CALCULATING VALUES OF ORDINATES



Data from page 763

of the given distribution; (2) that the standard deviation of the normal distribution shall be the same as that of the actual distribution; and (3) that the total frequency of the actual and the fitted distributions shall be the same. That is, N and σ are established by the actual distribution, and these are the only values needed in solving the normal-curve equation. The other terms in the equation, π and e , are constants, hence for all normal curves there is a constant relation between the distribution of N —the total frequencies or the area—and the number of standard deviations from the center. This relation is entirely independent of the values of x and Y in any particular distribution.

Because of this constant relation in its proportions, the fraction of the area of any normal curve that will lie between the center ordinate and an ordinate erected at any given number of standard deviations away from the center has been definitely established. The method by which these fractions of area have been determined is not explained in this text. All that is necessary is an understanding of how to read values in the Table of Areas, which is reproduced in Appendix F.

The total area under the curve is taken as unity. Then the area on either side of the center ordinate is $\frac{1}{2}$. Distances from the center ordinate are measured in units of the standard deviation, i.e., in units of $\frac{x}{\sigma}$.³ The table is arranged so that each value is the fraction of the total area lying between the center ordinate and the abscissa $\frac{x}{\sigma}$ recorded at the left of the table. Thus the area between the center ordinate and $\frac{x}{\sigma} = +1$ is .34134 or 34.13 per cent of the total area. The area between the center ordinate and $\frac{x}{\sigma} = -1$ is also 34 per cent of the total area. That is, the Area Table is prepared for half of the area of the curve but can be used for either half since the curve is symmetrical. Then a range of $\frac{x}{\sigma} = \pm 1$ includes 68.27 per cent of the total area. Likewise $\frac{x}{\sigma} = \pm 2$ includes $2 \times .47725$ or 95.45 per cent of the total area, and $\frac{x}{\sigma} = \pm 3$ includes $2 \times .49865$ or 99.73 per cent of the total area. These values were referred to in the discussion of dispersion in chapter XVIII, page 451, and Figure 64. It will now be clear that a range of three times the standard deviation on either side of the arithmetic average includes practically the entire area of the normal curve. This relationship is of primary importance in judging the reliability of samples, as described in the next chapter.

Fitting to a Given Distribution.—The fitting of a normal distribution to a given frequency distribution is of practical value when knowledge of the parent universe must be inferred from the actual distribution as a sample. The most convenient method is by the use of the Table of Areas. Table 158 illustrates the computations involved in fitting normal frequencies to the distribution of monthly costs of electric current. The values of the arithmetic average and standard deviation are obtained in the usual way in the upper half of the table. Column 5 lists the limits of the successive classes of the distribution. The deviations of the several class limits from the arithmetic average are given in column 6 and these deviations in units of the standard deviation are given in column 7. The values of column 8 are obtained from the Area Table. For instance, the lower limit of the first class interval is 2.85 standard deviations less than the average. From the table .49781 is the fraction of the total area lying between the ordinate at

³ In computing $\frac{x}{\sigma}$, x and σ must be measured in the same unit, but it need not be the width of the class interval, as in the computation of Y_0 .

TABLE 158

NORMAL DISTRIBUTION FITTED TO A FREQUENCY DISTRIBUTION OF MONTHLY COST OF ELECTRIC CURRENT OF A DOMESTIC CONSUMER FOR A TEN-YEAR PERIOD

MONTHLY COST	(1) <i>f</i>	(2) <i>d_s</i>	(3) <i>fd_s</i>	(4) <i>fd_s²</i>	$M' = 3.35$ $M = 3.35 + (\frac{2.3}{120} \times 1)$ $= 3.35 + .02 = 3.37$ $\sigma_s = \sqrt{\frac{4.87}{120} - (.19)^2}$ $= \sqrt{4.0583 - .0361}$ $= \sqrt{4.0222} = 2.005$ $\sigma = 2.005 \times .1 = .20$
\$2.80 - 2.89	1	-5	-5	25	
2.90 - 2.99	2	-4	-8	32	
3.00 - 3.09	6	-3	-18	54	
3.10 - 3.19	12	-2	-24	48	
3.20 - 3.29	22	-1	-22	22	
3.30 - 3.39	31	0	0	0	
3.40 - 3.49	19	1	19	19	
3.50 - 3.59	14	2	28	56	
3.60 - 3.69	5	3	15	45	
3.70 - 3.79	4	4	16	64	
3.80 - 3.89	2	5	10	50	
3.90 - 3.99	2	6	12	72	
	120		+23	487	

(5) CLASS LIMITS	(6) <i>x</i>	(7) <i>s</i> <i>σ</i>	(8) AREA BETWEEN CENTER ORDI- NATE AND $\frac{x}{\sigma}$	(9) FRACTION OF TOTAL AREA IN EACH CLASS	(10) FITTED FREQUEN- CIES (9) × 120
2.80	- .57	-2.85	- .49781	.00720	.86
2.90	- .47	-2.35	- .49061	.02277	2.73
3.00	-.37	-1.85	- .46784	.05635	6.76
3.10	- .27	-1.35	- .41149	.10915	13.10
3.20	- .17	-.85	- .30234	.16551	19.86
3.30	- .07	-.35	- .13683	.19645	23.57
3.40	+ .03	+ .15	+ .05962	.18253	21.90
3.50	+.13	+ .65	+ .24215	.13278	15.93
3.60	+ .23	+1.15	+ .37493	.07560	9.07
3.70	+ .33	+1.65	+ .45053	.03369	4.04
3.80	+ .43	+2.15	+ .48422	.01176	1.41
3.90	+ .53	+2.65	+ .49598	.00320	.38
4.00	+ .63	+3.15	+ .49918		
				.99699	119.61

$\frac{x}{\sigma} = 2.85$ and the center ordinate. All of the values in column 8 are read similarly from the table. Then the difference between two adjacent values in column 8 will give the fraction of the total area lying between the two class limits represented by those values. Thus the

fraction of the total area included in the class \$2.80 to \$2.90 is .49781 — .49061 = .00720 as shown in column 9. All of the values in column 9 are obtained in this way except the area between \$3.30 and \$3.40. Part of this area lies to the left of the center ordinate and part to the right. Therefore the fraction of the total area in this class is the *sum* of the area between \$3.30 and the center ordinate and the area between the center ordinate and \$3.40. As a guard against missing this shift in the direction, it is suggested that the algebraic signs of column 7 be repeated in column 8. Then the two parts of the interval whose limits are marked by the last minus sign and the first plus sign will always be added unless the true average happens to coincide with a class limit. In the latter case there is no class interval containing the center ordinate. However, the signs in column 8 have no algebraic significance, since these fractions of area represent frequencies and cannot be negative.

The sum of the fractional areas in column 9 should always be approximately equal to unity or just under unity, since there are no values fitted at the extreme limits of the normal curve. The fitted frequencies of column 10 corresponding to the actual frequencies of column 1 are obtained by multiplying the fractions of column 9 by 120, the total frequency of the given distribution. u

Figure 110 exhibits the original distribution, the fitted binomial, and the fitted normal distributions. The fitted normal curve is actually only a polygon obtained by joining values on the normal curve falling at the center of the class intervals of the given distribution. It appears to be a skewed curve but is not, because the trace of the symmetrical curve can be obtained by locating the center at the arithmetic average and at the height Y_0 and then connecting the plotted midpoints including this maximum ordinate. In fitting a normal distribution to a given set of data this procedure is unnecessary; accordingly reference was made earlier in the chapter to the greater flexibility of fitting the normal curve as compared with fitting the binomial.

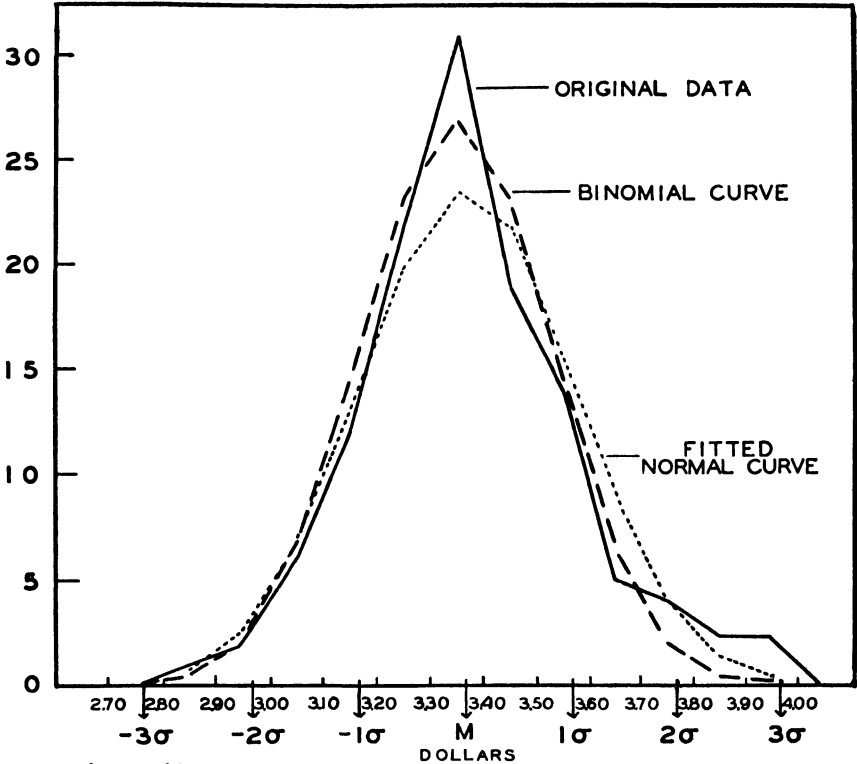
Neither of the fitted curves is a very close reproduction of the original. There is even some question as to whether the normal curve is a better fit than the binomial.⁴ This is precisely the sort of question

⁴ It must be remembered that in this comparison the actual distribution and the fitted normal curve have the same standard deviation, whereas the binomial has not been adjusted. Its standard deviation is 173 as compared with 200 for the actual and normal. If it were adjusted according to the method used in Figure 108, the peak of the binomial would be lower than that of the normal curve. This adjustment is not made in fitting a binomial because it is intended only as an approximation of the distribution of the universe.

FIGURE 110

BINOMIAL AND NORMAL DISTRIBUTIONS FITTED TO FREQUENCY DISTRIBUTION OF MONTHLY COST OF ELECTRIC CURRENT

NUMBER



Data from Tables 157 and 158

that arises in finding what type of curve to fit. A complete equipment of different kinds of curves is provided in advanced statistical work. The present treatment does not go beyond the binomial and the normal distributions. An exact way of judging which of these two types is the better fit to the distribution of electric-light bills will be given later in this chapter.

In many practical cases the fitted normal frequencies will provide useful information even though the given distribution does not conform very closely to the normal. In fact, one of the principal reasons for fitting the normal frequencies is to discover the extent of the departure of the actual from the normal. For this reason the use of the normal frequencies is not confined to distributions that are approximately in the symmetrical normal form.

Applications of the Table of Areas

In applied work the Table of Areas serves several purposes that are perhaps more important than the fitting of normal frequencies. These can be illustrated by presenting the solutions to several examples.⁵

Example I.—The cost accounting department of a wholesale drug concern had found that orders amounting to less than \$20 were usually filled at a loss. A study of the orders filled during one week showed that the average per order was \$38.16 with a standard deviation of \$12.32. Assuming that the distribution of orders is approximately normal, and that the sample is representative as to size of order, what percentage of all the orders will be filled at a loss?

$$x = \$20 - \$38.16 = -18.16$$

$$\frac{x}{\sigma} = \frac{-18.16}{12.32} = -1.47$$

The fraction of the total area of the normal curve between the center ordinate and the ordinate at $\frac{x}{\sigma} = -1.47$ is .42922. Therefore, over a period of time the drug company can expect that about 7 per cent of its orders will amount to less than \$20, because $.5 - .42922 = .07078$.

If the drug company were to institute a service charge on orders less than \$20, it could expect that the charge would be applicable at the outset to 7 per cent of the orders. The per cent would decline subsequently because some retailers would increase the size and decrease the frequency of their orders, while other small customers would cease buying from this company. The net loss of volume to the company would be slight and the net profit would probably be increased. The adverse effect on good-will might, however, outweigh all the advantages.

Example II.—An oil company has found from previous experience that its tank trucks have an average service life of 5.20 years with a standard deviation of 1.46 years. Assuming that previous experience is applicable to a new fleet of trucks and that the service life of trucks follows a normal distribution, how many out of a new fleet of 80 trucks will have to be replaced during the third year?

⁵ Although these examples parallel actual investigations, they are not sufficiently exact reproductions to use as sources of data.

This problem requires the fraction of the total area of the normal distribution falling in the two-year to three-year interval. Two years is —3.20 years from the center ordinate and three years is —2.20 years from the center ordinate. Expressing these deviations in units of the standard deviation and taking the areas from the table gives the following,

x	$\frac{x}{\sigma}$	Areas
— 3.20	— 3.20 ÷ 1.46 = — 2.19	.48574
— 2.20	— 2.20 ÷ 1.46 = — 1.51	.43448
		.05126

The fraction of area standing on the base 2–3 years is .05126. Therefore $80 \times .05126 = 4.1008$ or 4 trucks is the expected replacement during the third year after the purchase of the new trucks.

Example III.—A manufacturer had an order for 10,000 half-inch carriage bolts. Standard specifications on these bolts call for the heads to be .25 inches in thickness with a tolerance of .01 inches either way. A test of the thickness of the heads of the bolts showed a standard deviation of .0063 inches from the required thickness of .25 inches. On the basis of this test how many bolts of a run of 10,000 could be expected to fail to meet the specifications?

This problem requires finding the portion falling beyond .01 inches at each end of the normal distribution. The ordinate is,

$$\frac{x}{\sigma} = \frac{.01}{.0063} = 1.59$$

and the area beyond $\frac{x}{\sigma} = 1.59$ is $.5 - .44408 = .05592$. The area beyond .01 inches at the two ends of the distribution is $2 \times .05592 = .11184$. Hence $.11184 \times 10,000$ or 1118 bolts are likely to be rejected because the thickness of the heads is not according to the specifications. Having arrived at this conclusion two alternative lines of action may be followed. If other orders are being filled in which the specifications are less rigid, the rejects from this order can be used in other orders. If most orders specify a tolerance of .01 inches, the producing machinery should be adjusted to increase the precision of manufacture.

Example IV.—In Example III the thickness of the heads of how many bolts in an order of 10,000 should fall within .002 inches of the .25-inch standard?

This problem requires finding what fraction of the normal distribution falls within the limits, $M \pm \frac{x}{\sigma}$.

$$\frac{x}{\sigma} = \frac{.002}{.0063} = .317$$

The area between the center ordinate and $\frac{x}{\sigma} = .317$ is .12438 (using straight-line interpolation). Then the area between $\pm \frac{x}{\sigma} = .317$ is $2 \times .12438 = .24876$, and $10,000 \times .24876$ or 2488 bolts should be produced with head thickness between .248 inches and .252 inches.

OTHER TYPES OF DISTRIBUTIONS

Figure 57 on page 380 shows two symmetrical curves, neither of which is normal, two skewed bell-shaped curves, and two unusual types, the J-shaped and the U-shaped curves. The properties and equations of all of these curves are explained in advanced statistical texts.⁶ Although many of the distributions of applied statistics resemble some one of these forms rather than the normal form, we shall have to be content to rely on a knowledge of the normal distribution, because the more complete development lies outside the scope of this book and beyond the mathematical knowledge the reader of this book is presumed to possess.

The loss on this score is less vital than the preceding statement seems to imply. Many sample distributions that appear to depart materially from the normal form come from universes that are much closer to normal. The evidence on this point can be studied by reverting to the binomial distribution.

The binomial expansions discussed at the beginning of the chapter dealt exclusively with the form of the distribution of two characteristics that have an equal chance of occurrence, i.e., $p = q = \frac{1}{2}$. These expansions were symmetrical and as n increased they approached the normal form. When p and q are not equal, the binomial distribution produces a skewed curve. This property is shown in Figure 111-A which depicts the distributions of Table 159. As the value of p increases from .05 to .4 the successive curves approach closer to symmetry although an appreciable amount of skewness remains. The mode of the $p = .4$

⁶ See particularly, W. Palin Elderton, *Frequency Curves and Correlation* (Cambridge; England: Cambridge University Press, 1938).

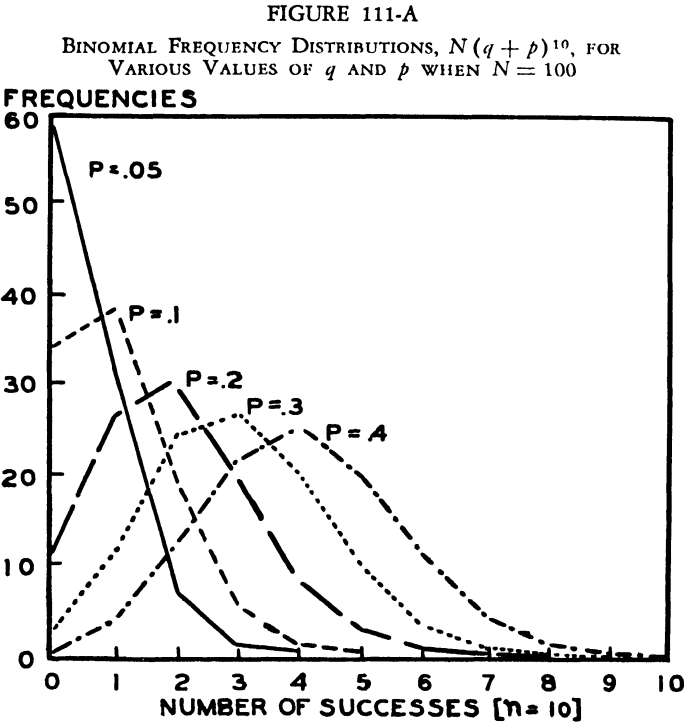


TABLE 159

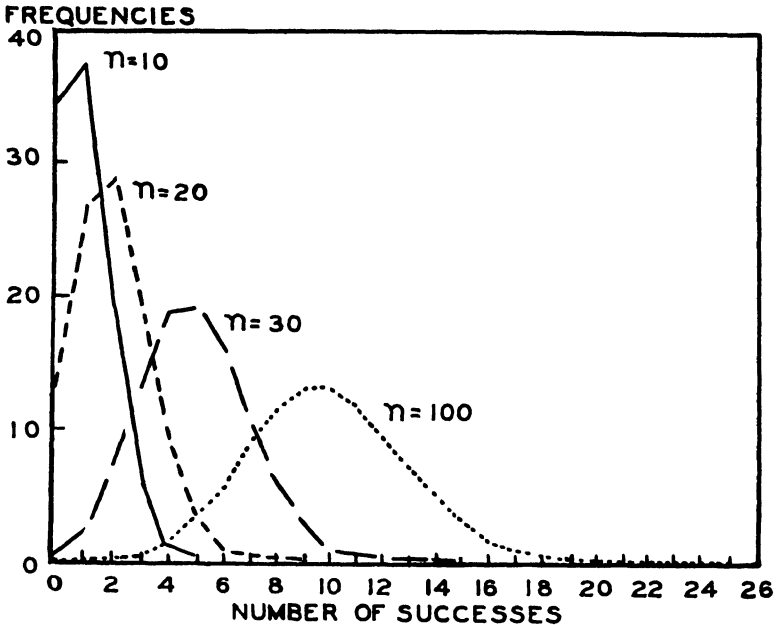
BINOMIAL FREQUENCY DISTRIBUTIONS, $N(q + p)^{10}$, FOR
VARIOUS VALUES OF q AND p WHEN $N = 100$

No. of Successes	$q = .95$ $p = .05$	$q = .9$ $p = .1$	$q = .8$ $p = .2$	$q = .7$ $p = .3$	$q = .6$ $p = .4$
0	59.87	34.87	10.74	2.82	.60
1	31.51	38.74	26.84	12.11	4.03
2	7.47	19.35	30.20	23.35	12.09
3	1.05	5.74	20.12	26.69	21.49
4	10	1.12	8.80	20.02	25.07
5		.15	2.64	10.30	20.06
6		..	.53	3.68	11.15
7		..	.07	.90	4.24
8			.01	.14	1.06
9				.01	.15
10					.01
	100.0	100.0	100.0	100.0	100.0

curve is in the fifth class interval, whereas in a normal distribution the mode would fall in the sixth class. An actual distribution skewed as much as $q = .7$, $p = .3$ could be analyzed approximately by the use of

FIGURE 111-B

BINOMIAL FREQUENCY DISTRIBUTIONS, $N(q + p)^n$, FOR
VARIOUS VALUES OF n WHEN $q = .9$, $p = .1$, AND $N = 100$



Data from Table 160.

the normal distribution; although a better fit could be obtained by more advanced methods, if a very precise analysis were required.

The tendency of skewed curves to approach the normal is further demonstrated in Table 160 and Figure 111-B. All of the curves in this graph are drawn from distributions of $100(.9 + .1)^n$. For $n = 10$ this distribution is decidedly skewed but as n is increased the skewness becomes less pronounced.

In Figure 111-B the distribution $100(.9 + .1)^{100}$ has 101 terms but those beyond the twenty-sixth have less than .001 per cent of the total frequency, hence are omitted from the diagram. Similar statements apply to the other distributions. The maximum ordinate of each of these curves is far away from the center of the range, at which point it would fall in the distribution $100(\frac{1}{2} + \frac{1}{2})^n$, but the cluster of the frequencies around the maximum ordinate resembles the cluster of the normal curve.

On an earlier page of the chapter it was pointed out that when $p = q = .5$ the symmetrical binomial approaches the normal as n is

TABLE 160
BINOMIAL FREQUENCY DISTRIBUTIONS, $N(q + p)^n$, FOR VARIOUS VALUES
OF n WHEN $q = .9$, $p = .1$, AND $N = 100$

No OF SUCCESSSES	BINOMIAL FREQUENCIES, $q = .9$, $p = .1$			
	$n = 10$	$n = 20$	$n = 50$	$n = 100$
0	34 87	12 16	52	003
1	38 74	27 02	2 86	030
2	19 35	28 52	7 79	.162
3	5.74	19 01	13 86	.589
4	1 12	8 98	18 09	1.588
5	.15	3 19	18 49	3 387
6		.89	15 41	5 958
7		.20	10 76	8 889
8	.	.04	6 13	11 483
9	.	01	3 33	13 043
10		1 52	1 52	13 188
11	.	.	61	11 989
12		.	22	9 879
13	.	.	07	7 431
14		.	02	5 131
15		.	01	3 268
16		...		1 929
17	.	.	.	1 059
18		...	...	.543
19		...	...	.260
20	.	.	.	.117
21			.	050
22		..	.	020
23		.	..	007
24		.	.	003
25		..		.001
	100 0	100 0	100 0	100 0

increased. The curves of Figure 111-B show that the same tendency toward normality with increasing n is present even though $p = .1$ and $q = .9$. In a continuous distribution the number of classes that can be used in studying a sample is small compared with the number that could be used in studying the entire universe, if the latter were available.⁷ It follows that a small value of n in a sample distribution may introduce an appearance of skewness not actually present in the universe. Conversely then we can expect that universe distributions are more nearly normal than sample distributions. For that reason normal-curve analysis can be employed in inferring conditions of a universe even though the sample that represents it is appreciably skewed.

The limiting case in which either p or q becomes exceedingly small has occasional applications in dealing with an event that occurs infre-

⁷ See chapter XV, p. 368.

quently but persistently, e.g., the chance that an error will be made in writing the checks for the monthly payroll of an industrial concern. Not many such errors are made but sometimes as many as five may creep into the writing of 800 checks in a given month, whereas in another month no errors may occur. A phenomenon of this kind is studied by means of the Poisson distribution which is an approximation of the binomial in the case where $q - p$ approaches unity.⁸

GOODNESS OF FIT—CHI-SQUARE TEST

The preceding discussion emphasized the use of normal-curve analysis even in distributions that depart materially from the precise normal form. Inasmuch as such wide application of the normal distribution is being suggested, the question of the conformity of the fitted distribution to a given set of data comes to have major importance. Hitherto in this book the test of conformity has been by inspection only. At this point an objective test of goodness of fit is introduced.

As applied to a frequency distribution the test consists in finding the square of the variation of each actual frequency from the corresponding fitted frequency, dividing each squared difference by the corresponding fitted frequency and summing these relatives for all classes of the distribution. This sum is known as chi-square (χ^2) and

may be expressed in symbols, as $\chi^2 = \sum \left\{ \frac{(f_a - f)^2}{f} \right\}$, in which

f_a = the actual frequencies

f = the frequencies of the fitted curve

If $(f_a - f)$ is zero for every class interval, the value of χ^2 is zero. If the differences between the two sets of frequencies are large compared with the fitted frequencies, the value of χ^2 will be large. In general, then, a large value of χ^2 denotes a poor correspondence between the actual and fitted distributions, while a small value of χ^2 denotes a good fit. However the actual value of χ^2 depends upon the number of class intervals in a given distribution; therefore its values are not direct criteria of goodness of fit. •

⁸ The Poisson distribution is explained in G. Udny Yule and M. G. Kendall, *An Introduction to the Theory of Statistics* (London: Charles Griffin and Co., Ltd, 1937), pp. 187-91.

The probability of values of χ^2 as a result of chance alone follow a definite law of distribution. Accordingly tables have been prepared showing the probability of the chance occurrence of any computed value of this measure. The tables are set up so that the probability of occurrence of any computed value of χ^2 can be determined for a distribution of given size.⁹

The testing of the correspondence between an actual distribution and a fitted distribution, as well as the testing of small samples in the next chapter, requires the introduction of a new concept known as the "degrees of freedom of the estimate." The theoretical basis for the determination of the degrees of freedom of an estimate cannot be brought into an elementary text, but the use of the concept can be explained by means of examples. If a fitted distribution could be established independently, that is, without any reference to the given distribution, then there would be no loss of freedom in the estimate. On the other hand each value from the given distribution used in establishing the fitted distribution constitutes a constraint upon the estimate, or the loss of a degree of freedom.

In testing by χ^2 the total possible number of degrees of freedom is limited by the number of values of $(f_a - f)$, but "the number of degrees of freedom must be reduced by unity for each constant of the universe [fitted distribution] which is estimated from the data [given distribution]." ¹⁰ In accord with this rule the fitted binomial loses one degree of freedom because the fitted distribution must have the same total frequency as the given distribution. The estimate is not free to vary in this respect. Another constraint usually arises from the assumption of values of p and q .¹¹

In fitting a normal distribution, three degrees of freedom are lost because the total frequency, the value of the arithmetic average, and the value of the standard deviation of the given distribution are used in computing the normal frequencies.

The notation used for degrees of freedom is,

N = the number of values of $(f_a - f)$ ¹²

m = the number of constraints or degrees of freedom lost

$N - m$ = the number of degrees of freedom of the estimate (χ^2)

⁹ Karl Pearson, *Tables for Biometricians and Statisticians* (London: Biometrika Office, University College).

¹⁰ Yule and Kendall, *op. cit.*, p. 428.

¹¹ This constraint is absent, however, in a controlled experiment in which the values of p and q are known from the universe.

¹² Note that in the χ^2 test N refers to class intervals and not to the total frequency.

Figure 112 contains all of the essential information that can be obtained from a table of χ^2 and can be used instead of a table. The first vertical line of the diagram represents the degrees of freedom ($N - m$). The second line represents the values of χ^2 and the third line represents the probability that under the given conditions the computed value of χ^2 will occur as a result of chance. To use the diagram lay a ruler at the proper values of $N - m$ and χ^2 ; the point at which the ruler intersects the P scale indicates the probability that the computed value of χ^2 could have occurred as a chance event. If the ruler falls on a high value of P the variation of the actual from the fitted distribution is highly probable. On the other hand, if the ruler falls on a low value of P , the variation is improbable as a chance occurrence and one may presume the presence of some significant cause of variation other than chance. There remains, however, the question of exactly when a given value of χ^2 should be considered significant. In common practice the 5 per cent level of probability is considered to be significant, that is, if a given $N - m$ and χ^2 lead to $P \leq .05$, the odds are 19 to 1 against the variation arising from chance and one may be justified in assuming that a significant cause of variation exists. In other words the result can be taken to indicate that the actual data cannot be represented by the fitted data.

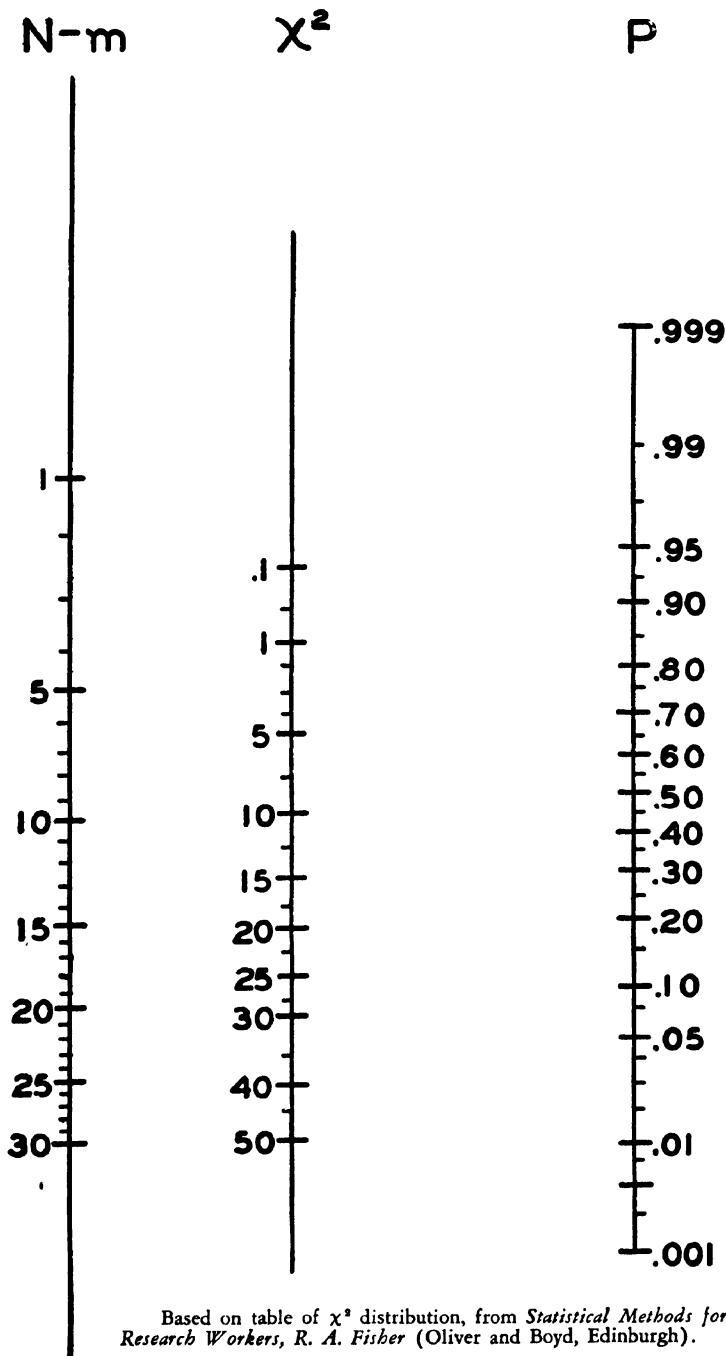
The 1 per cent level of significance is sometimes used when the observer prefers to be more conservative. In such cases the probability that a given variation arises from chance is .01 or the odds against it 99 to 1. One can therefore feel some confidence in saying that a value of $P \leq .01$ indicates the presence of a significant cause of variation. It is impossible to state a rule on this point, but it can be said in general that $.05 > P > .01$ strongly suggests significance and $P < .01$ warrants a presumption of significance.

The computation of χ^2 for the normal distribution fitted to monthly electric costs is shown in Table 161. The number of class intervals has been reduced from twelve to seven by combining those having a small number of frequencies. A general rule applies to this step: in computing χ^2 , no class interval should be included with less than *five* frequencies and a minimum of *ten* is preferable if the particular distribution will permit.

The value of χ^2 turns out to be 3.69. The question is whether this value indicates that the actual distribution differs significantly from the normal or not. Entering Figure 112 at $N - m = 4$, $\chi^2 = 3.69$

FIGURE 112

DIAGRAM FOR FINDING VALUES OF P ASSOCIATED WITH
COMPUTED VALUES OF χ^2 AND $N - m$



Based on table of χ^2 distribution, from *Statistical Methods for Research Workers*, R. A. Fisher (Oliver and Boyd, Edinburgh).

TABLE 161

COMPUTATION OF χ^2 OF A NORMAL DISTRIBUTION FITTED TO MONTHLY COST OF ELECTRIC CURRENT, DATA FROM TABLE 158

MONTHLY COST	f_a	f	$f_a - f$	$(f_a - f)^2$	$\frac{(f_a - f)^2}{f}$
Less than \$3.10 ...	9	10.35	-1.35	1.82	.18
3.10-3.19	12	13.10	-1.10	1.21	.09
3.20-3.29	22	19.86	+2.14	4.58	.23
3.30-3.39	31	23.57	+7.43	55.20	2.34
3.40-3.49	19	21.90	-2.90	8.41	.38
3.50-3.59	14	15.93	-1.93	3.72	.23
3.60 and over.....	13	14.90	-1.90	3.61	.24
	120	119.61			3.69

$$N = 7; m = 3; N - m = 4; \chi^2 = 3.69; P = .46$$

gives $P = .46$. Therefore a variation as great as this might occur in 46 out of 100 cases as a chance event and is certainly not to be considered as significant.

A question was raised earlier in the chapter concerning the relative merits of the normal frequencies and the binomial frequencies as a fit to the distribution of monthly electric-current costs. The χ^2 test is applied to the binomial frequencies in Table 162. The value of χ^2 turns out to be 3.90. There are two constraints in the binomial, the fact that the total frequency must be 120 and the fact that the particular fitted distribution $p = q = .5$ has been *assumed* as a description of the given distribution. Hence $N - m = 7 - 2 = 5$. In Figure 112 the probability of $\chi^2 = 3.90$ with 5 degrees of freedom is a little more than .60. That is, variability as great as this could be expected in six cases out of ten.

The test indicates that the binomial is a better fit than the normal

TABLE 162

COMPUTATION OF χ^2 OF A BINOMIAL DISTRIBUTION FITTED TO MONTHLY COST OF ELECTRIC CURRENT, DATA FROM TABLE 157

MONTHLY COST	f_a	f	$f_a - f$	$(f_a - f)^2$	$\frac{(f_a - f)^2}{f}$
Less than \$3.10....	9	8.76	+ .24	.0576	.01
3.10-3.19	12	14.50	-2.50	6.2500	.43
3.20-3.29	22	23.20	-1.20	1.4400	.06
3.30-3.39	31	27.07	+3.93	15.4449	.57
3.40-3.49	19	23.20	-4.20	17.6400	.76
3.50-3.59	14	14.50	-.50	.2500	.02
3.60 and over.....	13	8.76	+4.24	17.9776	2.05
					3.90

$$N = 7; m = 2; N - m = 5; \chi^2 = 3.90; P = .60$$

distribution. The values of χ^2 of both fitted distributions are highly probable. Study of the amounts contributed to χ^2 by the individual classes shows that the given distribution rises to a higher peak than either of the fitted distributions. The binomial in turn rises higher than the normal. Hence the normal is the poorer fit at the center as indicated by the large contribution of this class to the total value of χ^2 for the normal distribution. On the other hand, more than half of the total value of χ^2 in the binomial is contributed by the upper class, \$3.60 and over.

Investigation of the algebraic signs of the $(f_a - f)$ column of the normal computation shows that the frequencies of the given distribution are more concentrated than the frequencies of the normal distribution. But a similar survey of the binomial computation shows that the fitted and the actual are about equally concentrated. This difference in fit of the normal and the binomial is precisely what is expressed in the higher value of P for the binomial. The comparison of these two fitted distributions by means of the χ^2 test provides a good example of the amount of detailed information that can be obtained concerning a fitted distribution by studying not only the total value of χ^2 but also the amounts contributed to χ^2 by each class interval.

One would conclude that this sample of domestic electric costs conforms to both the binomial and the normal distributions well enough to warrant the use of either as a description of the way such costs will be distributed. The test proves nothing concerning the representativeness of the particular sample, but rather rests on the assumption that it is typical of the way electric costs are distributed. Several experiments similar to this one would be desirable as a basis for generalizing the conclusion stated. If the same correspondence of the actual and fitted distributions appeared repeatedly, the representativeness of the samples would be confirmed and the general applicability of the results would be indicated.

A word of caution is in order at this point. A value of P as low as .01 or .05 indicates the presence of a significant cause of variation. On the other hand a higher value of P does not prove that the fitted curve is a proper description of the given data. The high value indicates only that the hypothesis of the particular fitted distribution is not disproved. The question of whether some other type of distribution might be preferable as a description of the given distribution remains open. If another fitted distribution were tried and the χ^2 test led to

a markedly higher value of P , this fact would carry a fair presumption that the second hypothesis was a better representation. Even this reasoning, however, reaches a practical limit because recent advanced work demonstrates that values of $P > .99$ indicate that something other than chance is operative. That is to say, a fit as good as $P = .995$ would hardly ever occur as a chance event; hence very high values of P suggest a significant cause of lack of variation for the same reason that very small values of P suggest a significant cause of variation.

The test of goodness of fit is only one of the uses of the chi-square function. The others lie outside the scope of the present treatment.¹³

PROBLEMS

1. Explain the difference between "either or" and "both and" probability.
2. If two dice are rolled, what is the probability of a sum (a) of six? (b) less than nine? (c) equal to or greater than nine? (d) of seven or eleven? (e) greater than one? (f) of seven and eleven in two successive rolls of the two dice?
3. A classroom contains five rows of seven chairs each. If a certain class has 30 students, what is the probability that the fourth chair in the third row will be vacant? (Assume that the students occupy chairs purely at random).
4. Using the Pascal triangle on page 755 write the integral frequencies obtained from the expansion of $2^n(q + p)^n$ when $p = q = \frac{1}{2}$ and $n = 15$.
5. Check the results of Problem 4 by substitution in the binomial formula.
6. The distribution of orders of a wholesale drug company referred to in Example I, page 769, was as follows:

SIZE OF ORDER	NO. OF ORDERS
Less than \$10.	3
10 and less than 20	27
20 and less than 30	105
30 and less than 40.	223
40 and less than 50	133
50 and less than 60	64
60 and less than 70	19
70 and less than 80.	6
80 and less than 90.	2
90 and over.	8*
	592

* These 8 orders have been omitted from the total.

¹³ The student interested in pursuing this subject further will find ample material in Yule and Kendall, *op. cit.*, chapter 22.

- a) By the use of the Table of Areas find the fitted frequencies for this distribution.
 - b) Prepare a graph of the given and fitted distributions.
 - c) Is the assumption warranted that the orders of this drug wholesaler are approximately normally distributed? Explain.
7. Given the distribution of hourly wages in a steel plant.

HOURLY WAGES (in cents)	No. of WORKERS
40- 49.9	30
50- 59.9	215
60- 69.9	590
70- 79.9	1214
80- 89.9	730
90- 99.9	206
100-109.9	15
	3000

Assuming that this distribution of wages is typical of the wages of steel workers and that normal-curve analysis can be employed, consider the following questions for steel workers in general:

- a) If a minimum wage of 55 cents per hour were established, what percentage of the workers would benefit?
 - b) Is it true that only 10 per cent of steel men earn as much as 90 cents per hour?
 - c) If a wage increase were contemplated for workers earning less than 80 cents, what percentage of the labor force would receive increases?
 - d) What is the probability that the average wage for all workers is 75 cents per hour?
 - e) If another sample of wages of steel workers had an average of 72½ cents with a standard deviation of 10 cents, what are some possible explanations of the difference between the two samples?
8. The following is the age distribution of 3,124 passenger automobiles reported to the Iowa Bureau of Motor Vehicles as scrapped during the year 1922 in the state.
- Assume that the sample is representative of the ages at which passenger automobiles are scrapped in the United States and that the distribution is near enough to normal to warrant normal-curve analysis.
- a) What percentage of automobiles are likely to be operated for fifteen years or more? between fourteen and fifteen years?
 - b) What percentage are likely to be scrapped during the seventh year?
 - c) What percentage are likely to be scrapped within the first two years?
 - d) What percentage are likely to be scrapped when they are between four and eight years old?

AGE WHEN SCRAPPED (years)	PERCENTAGE OF AUTOMOBILES*
0- 1	2.2
1- 2	2.6
2- 3	3.6
3- 4	6.0
4- 5	11.4
5- 6	16.6
6- 7	17.9
7- 8	17.0
8- 9	12.4
9-10	5.7
10-11	2.7
11-12	1.2
12-13	.5
13-14	.2
	100.0

* Robley Winfrey and Edwin B. Kurtz, "Life Characteristics of Physical Property," *Bulletin of the Iowa Engineering Experiment Station*, Vol. XXX (June 17, 1931), p. 65.

9. a) Test the goodness of fit of the normal frequencies obtained in Problem 6 (a) by χ^2 .
- b) What is the probability that the assumption is warranted that the orders of this drug wholesaler are normally distributed?
- c) What further assumptions would be necessary in order to infer from the fitted frequencies the distribution of orders of all wholesalers of drugs?
10. a) Determine binomial frequencies for the distribution of Problem 6.
- b) Compute χ^2 and test the goodness of fit of the binomial frequencies.
- c) According to the χ^2 test, which set of fitted frequencies would you prefer as a description of the distribution of the orders of this concern?
11. By the use of χ^2 determine whether any of the four samples in Table 23, page 786, of chapter XXIX differs significantly from the universe.

REFERENCES

BURGESS, ROBERT W., *Introduction to the Mathematics of Statistics*. Boston: Houghton Mifflin Co., 1927.

An explanation of the relation between the binomial and normal distributions.

ELDERTON, W. PALIN, *Frequency Curves and Correlation*. Cambridge, England: Cambridge University Press, 1938.

Chapters V and VI contain a detailed explanation of the Pearsonian system of curves. This reference will be most useful to students with considerable mathematical training.

FISHER, R. A., and YATES, F., *Statistical Tables for Biological, Agricultural, and Medical Research*. Edinburgh: Oliver and Boyd, 1938.

KENNEY, JOHN F., *Mathematics of Statistics*. New York: D. Van Nostrand Co., Inc., 1939.

Chapter I of Part II contains a somewhat advanced explanation of probability, binomial distribution, and normal distribution.

PEARL, RAYMOND, *Medical Biometry and Statistics*. Philadelphia: W. B. Saunders Co., 1923.

Chapter XI contains a very lucid statement of the relation of the binomial to the normal distribution.

YULE, G. UDNY, and KENDALL, M. G., *An Introduction to the Theory of Statistics*. London: Charles Griffin and Co., Ltd., 1937.

Chapter 10 contains a discussion of the binomial, normal and Poisson distributions. The χ^2 test is developed in chapter 22. A diagram giving the values of P corresponding to values of χ^2 faces page 540.

CHAPTER XXIX

PRINCIPLES OF SAMPLING AND TESTS OF SIGNIFICANCE

INTRODUCTION

THE distinction between census and sample data and the methods of obtaining representative samples were discussed in chapters IV and V. The detailed explanation of the relation of a sample to the universe from which it is taken and the measuring of the reliability of samples have been purposely deferred to this chapter. The postponement was necessary because of the close connection of sampling measures with normal-curve analysis.

A large part of practical statistical work depends upon the use of samples. Examples are price indexes, wage studies, time and motion studies, indexes of production and general business conditions, actuarial tables, and all sorts of tests of manufactured goods. Depending on the problem involved, a sample may be a list of prices, a life-history of 100,000 individuals, or a quart measure of prepared cow feed. The variety of uses of sampling in business is sufficient evidence of the practical importance of the subject.

THE BASIS OF SAMPLING

The Principle of Statistical Regularity

The operation of this principle has been explained and illustrated in chapter V. We begin here with the understanding that a reasonably large group of items selected at random from a much larger universe will exhibit the characteristics of the universe. The chief problem of sampling is to discover with what fidelity the characteristics of the universe will be exhibited.

Shape of Universe and Sample

Many controlled experiments have been carried out to discover the relation between the distribution of cases in the universe and in the sample. The results show that whatever the shape of the parent distribution, or universe, that shape will be approximately reproduced in the sample, assuming of course that proper care has been exercised

to preserve the representative character of the sample. One of the best of these experiments, conducted by Dr. W. A. Shewhart, is described as follows:

Let us start our study of sampling with an experiment in which 4,000 drawings of a chip from a bowl were made with replacement; that is, after drawing a chip, it was replaced and thoroughly mixed with the others before another was drawn.

TABLE 22*—MARKING ON 998 CHIPS FOR SAMPLING EXPERIMENT

MARKING ON CHIP X	NUMBER OF CHIPS	MARKING ON CHIP X	NUMBER OF CHIPS	MARKING ON CHIP X	NUMBER OF CHIPS	MARKING ON CHIP X	NUMBER OF CHIPS
— 3.0 . .	1	— 1.5 . . .	13	0 0 . .	40	1 5 . .	13
— 2.9	1	— 1.4	15	0 1 . .	40	1 6	11
— 2.8 . .	1	— 1.3	17	0 2 . .	39	1 7 . . .	9
— 2.7	1	— 1.2	19	0 3	38	1 8 . . .	8
— 2.6 . .	1	— 1.1	22	0 4 . .	37	1 9 . . .	7
— 2.5 . .	2	— 1.0	24	0 5	35	2 0	5
— 2.4 . .	2	— 0.9	27	0 6 . . .	33	2 1 . . .	4
— 2.3 . . .	3	— 0.8	29	0 7 . . .	31	2 2 . . .	4
— 2.2 . . .	4	— 0.7	31	0 8 . . .	29	2 3	3
— 2.1 . . .	4	— 0.6	33	0 9	27	2 4	2
— 2.0	5	— 0.5	35	1 0	24	2 5	2
— 1.9 . . .	7	— 0.4	37	1 1	22	2 6	1
— 1.8 . . .	8	— 0.3	38	1 2 . . .	19	2 7	1
— 1.7	9	— 0.2	39	1 3	17	2 8	1
— 1.6	11	— 0.1	40	1 4	15	2 9	1
						3 0	1

* [This is the table number in the source.]

TABLE 23*—GROUPED FREQUENCY DISTRIBUTIONS IN SAMPLING EXPERIMENT

CELL MIDPOINT	DISTRIBUTION IN BOWL	OBSERVED DISTRIBUTIONS			
		SAMPLE No 1	SAMPLE No 2	SAMPLE No 3	SAMPLE No 4
— 3.0	3	5	1	2	2
— 2.5	9	9	14	10	9
— 2.0	28	36	24	29	25
— 1.5	65	55	51	72	49
— 1.0	121	123	113	124	112
— 0.5	174	165	187	181	191
0	198	203	195	180	204
0.5	174	172	176	169	182
1.0	121	123	125	120	123
1.5	65	68	71	67	64
2.0	28	31	31	32	25
2.5	9	8	8	11	12
3.0	3	2	4	3	2

* [This is the table number in the source.]

In the bowl there were 998 circular chips on each of which there was a number. Forty chips were marked 0, 40 were marked —0.1, 40 were marked

+0.1, and so on as shown in Table 22. Before replacing a chip in the bowl, the number was recorded.

In this experiment we have as near an approach as is likely feasible to the condition in which the law of large numbers applies since, to the best of our knowledge, the same essential conditions can be maintained. The differences between successive numbers drawn are beyond our control.

Dividing the observed values into four sets of 1,000 each, we get the four grouped frequency distributions of columns 3, 4, 5, and 6 in Table 23. Column 2 gives the corresponding distribution in the bowl.

As is to be expected, no two of the observed distributions are the same, and no one of them is the same as that in the bowl.¹

But all of them exhibit the characteristics of the universe.² The averages of the four samples differ little from the average of the universe, and the same can be said concerning other measures such as the standard deviation.

A sample tends also to reproduce the shape of a non-normal universe. But in this case the unbalanced shape of the universe may be somewhat exaggerated in the sample, if the latter is small compared with the universe. This phenomenon can be expected from the characteristics of the binomial expansion studied in the preceding chapter. It will be recalled that the skewed expansion $(.9 + .1)^n$ tended toward the symmetrical form as n increased.

The characteristics observed in controlled experiments have also been found to hold about as well in examinations of business data, although the opportunities for obtaining repeated samples are limited in practice. The case is clear enough, however, to justify the general conclusion that in dealing with actual data a properly prepared sample will reproduce the shape and characteristics of the universe.

Distribution of a Large Number of Samples From the Same Universe

If a large number of samples is taken from the same universe, all of the samples will have similar characteristics but they will not be identical. Chance variability will be present. Consider the arithmetic averages of the samples. They will vary around the unknown value of the average of the universe. Some of the individual averages will be less than the true (universe) average, others will be greater. Small

¹ W. A. Shewhart *Economic Control of Quality of Manufactured Product* (New York: D. Van Nostrand Co., Inc., 1931), pp. 164-66

² The discs in the bowl actually represent an infinite universe in the sense that at any individual drawing the choices are so great, i.e., 1 out of 998.

deviations from the true average will occur more frequently than large deviations, but a few cases of extremely large deviations are likely to be found. Specifically the individual averages will be arranged in the *normal form*, and the average of these sample averages is the theoretically best estimate of the value of the average of the universe. The same argument leads to the conclusion that not only the arithmetic average but other measures, such as the median, quartiles, or standard deviation, computed from the samples will each form a normal distribution around its own average value. Any such distribution of the values of a measure in a large number of samples is known as a *sampling distribution*.

The normal form of the sampling distribution of a measure will prevail whether the parent universe is normal or has some other form. The persistence of this quality of normality makes it possible to abandon the clumsy and impractical use of a large number of parallel samples in favor of a single sample.

The Use of a Single Sample

In most practical uses of sampling the statistician is fortunate to have available one good sample. The cases in which multiple sampling is possible are extremely rare. Consider for example the difficulties involved if the Bureau of Labor Statistics were to undertake to obtain 50 or 100 samples of factory employment monthly in order to take the average of the changes in the several samples as the best estimate of the actual change occurring in all factory employment. The labor of collecting and analyzing these samples would be prohibitive, but an even greater objection is the fact that it has taken years of effort for the Bureau to establish a reporting service comprehensive enough to provide *one* representative sample from which to measure changes in factory employment. This example is typical of what would be found in sampling any economic phenomenon.

To avoid the requirement of a large number of samples, a statistical technique has been developed by which the characteristics of a universe can be inferred from those found in a single sample. This process of inference is based on the fact that a large number of equally good measurements of a characteristic will distribute themselves in the normal form.

Probability of Reproducing the Universe.—The Shewhart tests referred to on page 787 indicate that the shape of a single sample will

reproduce the general contours of the universe from which it is taken. But we need to know how close this conformity will be. That is, will the several measures of the sample equal those of the universe? Take the arithmetic average as an illustration. Experiments with controlled data show that the averages of a large number of samples will range themselves in normal form around the average of the universe. The same type of grouping can be expected with uncontrolled data.

A given sample is merely one of the large number of samples which, if they existed, would have averages arranged in normal form around the average of the universe. Hence the average of the given sample may fall anywhere within the range of the sampling distribution. This is not the limit of knowledge of the average of the given sample, because we possess considerable information about the probability of occurrence of the various frequencies of a normal distribution. Therefore by dealing with probabilities within limits instead of exact criteria, additional description of the universe can be gained from a single sample.

The probability that a given average will fall *exactly* at the center of the distribution of averages of samples is extremely small, but the probability that it will fall within a stated range on either side of the center of this distribution depends upon the fraction of the total area of the normal distribution included between the stated limits. While these statements have been made in connection with the arithmetic average, they are equally applicable to any other measure of a frequency distribution such as the median, quartiles, average deviation, or standard deviation.

Particular attention is directed to the fact that the normal distribution referred to here is the distribution of the *averages* of a large number of possible samples. The standard deviation of this sampling distribution differs from the standard deviation of a given sample. To avoid ambiguity the standard deviation of a potential sampling distribution will be referred to as a *standard error* and will be written with a subscript to denote the measure involved. As fixed notation then,

σ_M will mean the standard error of the arithmetic average

σ_{M_e} will mean the standard error of the median

σ_σ will mean the standard error of the standard deviation

etc.

Use of the Standard Error of a Single Sample.—The variability in a sampling distribution of a given measure is usually determined by

its standard error, that is, the standard deviation of the sampling distribution of that measure. The discussion of the preceding chapter indicated that 68 per cent of the area of a normal curve lies between the limits $M \pm \sigma$. Since a sampling distribution takes normal form it is to be expected that in 68 per cent of the samples the value of the measure will fall within the range of one standard deviation in either direction from the center ordinate of the sampling distribution. This is expressed as $\pm$ one standard error of the measure involved.

In practice multiple sampling is impossible; therefore this principle of distribution is applied to express the probability of occurrence of a particular value of a given measure in a single sample. For example, the chances are 68 out of 100 that the *average* of the one sample taken does not differ from the center ordinate of the sampling distribution of averages by more than $\pm \sigma_M$. In fact the most probable position of the average of the single sample is near the center of the sampling distribution. If it is near the center, its value is close to the value of the average of the universe. We start, therefore, by assuming that the best estimate of the value of the average in the universe is the value of the average in the sample. But the likelihood that the average of the universe will differ from the value assumed from the sample must be stated explicitly.

Since the true average of the universe $\pm \sigma_M$ can be expected to include 68 per cent of the sampling distribution of averages, it can be assumed that, in approximately two out of three cases, the average of a sample will be within a range of $\pm \sigma_M$ from the average of the universe. Likewise, the true average $\pm 2 \sigma_M$ can be expected to include 95.5 per cent of the sampling distribution; therefore in approximately 19 out of 20 cases the average of a sample can be assumed to fall within a range of $\pm 2 \sigma_M$ from the average of the universe. And the true average $\pm 3 \sigma_M$ can be expected to include 99.73 per cent of the sampling distribution; therefore in approximately 369 out of 370 cases the average of a sample can be assumed to fall within a range of $\pm 3 \sigma_M$ from the average of the universe.

These relations have been stated for the mean. They apply to all of the measures of a frequency distribution without change except for the substitution of the other measure wherever the word "average" or "M" now appears.

An event that can be expected to occur once in 370 trials is rather

improbable; therefore custom has led to the use of ± 3 standard errors as the limits of variation. But the probability of occurrence of any variation from any measure in a sample can be found in the Table of Areas. This probability (P) provides information in addition to the limiting values of any measure at ± 3 standard errors.

Measures of Standard Error.—The preceding discussion has tacitly assumed that the standard error of a sampling distribution of any measure can be computed without detailed knowledge of a large number of samples. This is in fact the case because in developing formulas for standard errors of the various measures it is necessary only to know that the sampling distributions would be normal or substantially so. These formulas are expressed in terms of the standard deviation of a single sample³ and the number of cases in the sample. This is in accord with the emphasis of chapter XVIII on the standard deviation as a tool in more advanced analysis. The proofs of the several formulas lie beyond the scope of this chapter. We shall, therefore, be content to state without proof the formulas⁴ that are most commonly used.

Standard error of the arithmetic average

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \quad (1)$$

Standard error of the median⁵

$$\sigma_{Me} = 1.25 \frac{\sigma}{\sqrt{N}} \quad (2)$$

Standard error of the per cent occurrence of an event

$$\sigma_p = \sqrt{\frac{pq}{N}} \quad (3)$$

in which p = the per cent occurrence of the event and $q = 1 - p$.

³ The mathematical development of these formulas is carried out with the standard deviation of the universe. But this standard deviation is usually unknown, hence we are forced to use the standard deviation of the sample as the best available estimate of the standard deviation of the universe.

⁴ There is actually a loss of one degree of freedom in using σ computed from the sample as an estimate of the standard deviation of the universe. Hence σ should really be

computed by the formula $\sqrt{\frac{\sum v^2}{N-1}}$ in all of these except formula 3. This adjustment has

so little effect in the realm of large samples that it can be neglected. Later in the chapter the adjustment will be introduced in the formulas for testing small samples.

⁵ The coefficients of formulas 2 and 4 are obtained by solutions using the Table of Areas; consequently they will give satisfactory results only in case the sample is substantially a normal distribution.

Standard error of the first or third quartile

$$\sigma_{q_1} = \sigma_{q_3} = 1.36 \frac{\sigma}{\sqrt{N}} \quad (4)$$

Standard error of the standard deviation

$$\sigma_{\sigma} = \frac{\sigma}{\sqrt{2N}} \quad (5)$$

Standard error of the Pearsonian coefficient of correlation

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}} \quad (6)$$

Standard error of the rank difference coefficient of correlation

$$\sigma_{\rho} = \frac{1}{\sqrt{N - 1}} \quad (7)$$

Relation Between Sample and Universe

The basic use of the standard error of a measure in a single sample as a means of describing the unknown value of the measure in the universe from which the sample was taken has been explained in preceding pages. The maximum that can be done in such a case is to assume that the value of the measure in the sample is the most probable value in the universe, and to add the probability that the true value will vary from the probable value only within measurable limits. For example, as this is being written (February, 1941), local draft boards are rejecting about 32 per cent of the men examined under the Selective Training and Service Act of 1940. The question raised is, Will this rate of rejections be maintained as the training program expands? The statistician would propose the question in somewhat different form. Assuming that draft-board experience to date is representative of what can be expected for the entire conscription program, what are the probable limits of variation in applying existing experience to future examinations of draftees? The pertinent data are,

Inducted—approximately 20,000 men⁶

Rejected—32 per cent

Examined— $20,000 \div .68 = 30,000$ approximately⁶

⁶ These figures are used in rounded numbers because there is no way of knowing the exact number of medical examinations from which the 32 per cent is computed. The percentage is believed to be reasonably accurate, however, because it originates from a medical source.

From formula 3, page 791, the standard error of sampling of a per cent occurrence is,

$$\sigma_p = \sqrt{\frac{pq}{N}}$$

In this illustration $p = .32$, $q = 1 - p = .68$, and $N = 30,000$. Therefore $\sigma_p = \sqrt{\frac{.32 \times .68}{30,000}} = .0027$ or .27 per cent. Since the limits of $\pm 3\sigma_p$ include nearly all of the probabilities, it is to be expected that rejections in the future will fall somewhere between 31.19 per cent and 32.81 per cent of the men examined.

The preceding sentence is based on the assumption that experience to date is representative of what will happen in future months. Specifically it means that there is no reason to believe that the men examined are either better or worse physically than those yet to be examined; that there is no expectation that examining physicians will either relax or tighten-up in their work; that draft executives will issue no general orders that will affect medical examinations.

Examples of this sort occur less frequently than those in which either information is available concerning the value of measures in the universe or, in the absence of definite knowledge, some hypothesis can be made or some standard established. When any such assistance is available, the sampling analysis can be carried much further. For example, sample collections of data concerning retail trade can often be compared with results of the Census of Business. A sample might be taken in a community to test the statement often made by writers and speakers that industry will not hire men over 40 years of age. A sample of a shipment of galvanized sheets might be tested to find out whether they fulfilled the specifications as to thickness of galvanized coating. Beyond these cases are others in which two samples are taken from the same universe and although nothing may be known concerning the values of measures of the universe, the two samples can be compared to obtain more information than would be available from either sample alone.

In all work of this type the major purpose is to determine whether the variations of the measures of the sample from the corresponding measures of the universe or other standards of comparison might have occurred as chance events or whether some non-chance factor has produced a significant variation. This subject is so important that

the remainder of the chapter is devoted to it. Tests of significance are set up in terms of the standard error of sampling, or a modification of it in the case of small samples. What follows is therefore the major application of the use of the standard error of a single sample as a measure of reliability, and of its counterpart, significance.

TESTS OF SIGNIFICANCE

When measures of a sample are to be compared with corresponding measures in the universe, the probability (P) that the difference is attributable to chance variation can be obtained from the Table of Areas, and levels of P can be established beyond which a difference is considered to be significant. A difference is said to be significant if the probability of its occurrence as a chance variation is so small that a hypothesis of the existence of non-chance factors is more tenable. The common practice is to suspect significant causes when $P < .05$, to search actively for such causes when $P < .01$, and to call values of $P < .001$ almost certain evidence of a significant difference.

Non-chance factors may be: the presence of some bias in taking the sample, failure to provide in the sample for all of the characteristics of the universe, an inadequate number of cases in the sample, or some concealed change in the universe during the sampling. These possible causes of significant differences are merely the respects in which a sample may fail to be representative of the universe from which it is taken. In detecting significant differences, the statistician may not be capable of indicating the specific cause but merely of recording the presence of such causes.

The uses of tests of significance can be explained by the introduction of several examples. The first group of examples pertains to some basic measures, the second to differences of basic measures, and the third to coefficients of correlation.

Reliability of Basic Measures

Example 1 (Arithmetic Average).—The theoretical thickness of 16-ounce copper sheets is .0216 inches. A concern manufacturing these sheets made a test of one hundred sheets and found that the thickness was as follows:

GAUGE THICKNESS IN INCHES	NO. OF SHEETS	
.0206 — .0208	4	$M = .02140$
.0208 — .0210	11	$\sigma = .00033$
.0210 — .0212	14	
.0212 — .0214	17	$\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{.00033}{10} = .000033$
.0214 — .0216	20	
.0216 — .0218	26	$\frac{\text{observed difference}}{\sigma_M} = \frac{.0216 - .0214}{.000033} = 6$
.0218 — .0220	7	
.0220 — .0222	1	
	100	

The difference between the observed mean and the theoretical mean is,

$$.0216 - .0214 = .0002$$

This difference divided by the standard error of sampling σ_M shows by how many standard errors the observed mean deviates from the theoretical mean. Thus, $\frac{.0002}{.000033} = 6$, and the observed mean differs from the theoretical mean by six times the standard error of sampling. A variation $\frac{x}{\sigma} = 6$ has a probability so small that it is not recorded in the Table of Areas. This manufacturer's product is significantly thinner than the standard. Unless he adjusts his machinery he runs an extraordinary risk of turning out a product that will fail to meet purchasers' specifications.

Example II (Arithmetic Average).—A research organization collected data for the year 1938 on sales and related information from 519 independent drugstores located in all parts of the United States. Omitting stores with annual sales of less than \$10,000, the sample was compared with the results of the 1935 Census of Business for the same type and size of stores. The average sales per store in the Census of Business was \$24,700. For the sample of 519 stores the average sales was \$32,500. Could this difference between sample and standard be due to chance? The standard deviation of the sample was \$16,700. Then

$$\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{\$16,700}{\sqrt{519}} = \$730$$

$$\frac{\text{observed difference}}{\sigma_M} = \frac{\$32,500 - \$24,700}{\$730} = 10.7 \quad \bullet$$

⁷ For all of the measures of significance of large samples the ratio, observed difference divided by standard error of sampling, will be referred to as $\frac{x}{\sigma}$ to conform to the notation of the Table of Areas.

The probability that a variation equal to 10.7 times the standard error of sampling could occur as a chance event is too small to record in the Table of Areas. The research organization's sample is not representative of all drugstores with respect to size of store as measured by sales.

Example III (Standard Deviation).—Cold-drawn seamless steel tubing with outside diameter of $5\frac{1}{2}$ inches is sold to a specification equivalent to a standard deviation of .020 inches. Thirty-three measurement tests on a carload of pipe show a standard deviation of .024 inches. What is the probability that the difference between the standard deviation of this sample and the specified variation is due to chance?

The standard error of sampling for the standard deviation according to formula 5, page 792, is,

$$\sigma_{\sigma} = \frac{\sigma}{\sqrt{2N}} = \frac{.024}{\sqrt{66}} = \frac{.024}{8.124} = .00295$$

$$\frac{\text{observed difference}}{\sigma_{\sigma}} = \frac{.024 - .02}{.00295} = \frac{+.004}{.00295} = +1.36$$

The probability of $\frac{x}{\sigma} = +1.36$ is, from the Table of Areas, $.5 - .4131 = .0869$. The range of standard error below .02 inches is of no importance in this test; hence the corresponding probability of variations at the negative end of the Area Table is disregarded. A standard deviation as great as .024 inches can be expected in 9 out of 100 samples. Therefore the test does not show any significant lack of uniformity in the outside diameter of this carload of tubing.

Example IV (Per Cent Occurrence).—A team of radio entertainers who had been broadcasting locally wished the management of a large department store to sponsor its program. The entertainers claimed that 20 per cent of the families in the local area were regular listeners to their program. Before signing a contract the department store decided to take a sample poll of listeners while the program was on the air. Two hundred phone calls were made, from which 30 families were reported as listening to the program in question. Was there a significant difference between the 15 per cent shown by the sample and the 20 per cent claim? According to formula 3, page 791:

$$\sigma_p = \sqrt{\frac{pq}{N}}$$

Then

$$\sigma_p = \sqrt{\frac{.15 \times .85}{200}} = .0252$$

$$\frac{\text{observed difference}}{\sigma_p} = \frac{.15 - .20}{.0252} = \frac{-.05}{.0252} = -1.99$$

and

$$P = (.5 - .47670) = .0233.$$

Therefore the chance is only 2 out of 100 that the entertainers' popularity was as great as they claimed.

If the prospective sponsors were interested in variations in *either* direction from 20 per cent as great as $\frac{x}{\sigma} = 1.99$, $2P$ would be used, giving a probability of 5 in 100.

Example V (Per Cent Occurrence).—Results over a long period of time showed that in a certain factory the operators of bolt-threading machinery were spoiling about 50 bolts per 1,000. Complaints of inspectors led to an investigation which showed that in a continuous run of 12,500 bolts 706 were spoiled. Is this a significant variation?

The expected spoilage was 5 per cent.

The observed spoilage was 5.65 per cent.

$$\sigma_p = \sqrt{\frac{pq}{N}} = \sqrt{\frac{.05 \times .95}{12,500}} = \sqrt{.0000038} = .00195$$

$$\frac{\text{observed difference}}{\sigma_p} = \frac{.0565 - .05}{.00195} = \frac{+.0065}{.00195} = +3.3$$

and

$$P = .5 - .49952 = .00048$$

This variation has such a slight probability of occurrence as a chance event that the management is justified in assuming that a significant cause is present. This cause might be inferior raw material, poor adjustment of machinery, inferior machine tools, inexperienced labor, or other factors. But an investigation of the cause or causes of the difficulty should be inaugurated immediately.

In Example IV the values of p and q observed in the sample were used in computing σ_p . In Example V p and q were taken from the universe. The rule on this point is: take the values of p and q from

the universe when they are available; use the p and q of the sample as the best available estimate of their values in the universe when no direct knowledge of the universe is at hand. In particular the 20 per cent estimate of the entertainers in Example IV would not be considered as reliable as the value of p obtained from the sample. But the long-term experience of 5 per cent spoilage in bolt threading is a reliable estimate from a large universe and should be used in computing σ_p instead of the value 5.65 per cent observed in the sample.

Reliability of Differences

A second use of standard error in testing reliability arises in comparing two samples to determine whether they differ significantly. The comparison may involve any of the basic measures such as the arithmetic average, standard deviation, or the per cent of occurrence.

If a large number of pairs of samples are taken from the same universe, the differences in values of any measure will distribute themselves in approximately normal form. Consider for example the arithmetic average and let M_1 stand for the average of the first sample of a pair and M_2 for the average of the second sample of the pair. Although these two averages should not differ by much, we know from the previous discussion that they are not likely to coincide exactly. Then if a second subscript is used to denote successive pairs of samples the differences (d) between a large number (n) of such pairs can be described symbolically as follows:

$$M_{11} - M_{21} = d_1$$

$$M_{12} - M_{22} = d_2$$

$$M_{13} - M_{23} = d_3$$

$$M_{14} - M_{24} = d_4$$

$$\cdot$$

$$\cdot$$

$$\cdot$$

$$M_{1n} - M_{2n} = d_n$$

Some of the d 's will be positive and others negative, their average value will be approximately zero, and small differences (i.e., small deviations from zero) will occur more frequently than large ones. In short they will be arranged in the normal form. This is a sampling distribution of the same type as those previously discussed. Its standard deviation is also a standard error of sampling. The values of this

The usual computations give

$$\begin{array}{ll} M_1 = 10.7967 & \sigma_1 = .246 \\ M_2 = 10.7233 & \sigma_2 = .334 \end{array}$$

Then by formula 8, page 799,

$$\sigma_{M_1 - M_2} = \sqrt{\frac{(.246)^2}{60} + \frac{(.334)^2}{60}} = \sqrt{\frac{.0605 + .1116}{60}} = \sqrt{\frac{.1721}{60}} = .0536$$

$$\frac{\text{observed difference}}{\sigma_{M_1 - M_2}} = \frac{10.7967 - 10.7233}{.0536} = \frac{.0734}{.0536} = \pm 1.37$$

$$P = .5 - .4147 = .0853$$

and

$$2P = .1706$$

Therefore a difference in average weight per lineal foot of at least .0734 pounds could be expected as a chance occurrence in 17 out of each 100 pairs of samples. According to this test there is no significant weight difference between the two car loads of bars. $2P$ is used here because differences as great as $\frac{x}{\sigma} = 1.37$ in either direction are equally important.

Example VII (Difference of Standard Deviations).—The shapes of the sample distributions in the preceding example suggest that there may be a significant difference in the uniformity of the weights of the bars in the two car loads. This could be investigated by comparing the difference between the two observed standard deviations with the appropriate standard error, namely, that for differences between standard deviations of pairs of samples. Then $\frac{x}{\sigma}$ would be the ratio of the difference between the two observed standard deviations to this standard error.

From Example VI,

$$\sigma_1 = .246 \qquad \sigma_2 = .334 \qquad 2N_1 = 2N_2 = 120$$

Substituting these values in formula 9, page 799,

$$\begin{aligned} \sigma_{\sigma_1 - \sigma_2} &= \sqrt{\frac{(.246)^2}{120} + \frac{(.334)^2}{120}} = .038 \\ \frac{\text{observed difference}}{\sigma_{\sigma_1 - \sigma_2}} &= \frac{.334 - .246}{.038} = \frac{.088}{.038} = \pm 2.32 \end{aligned}$$

$$P = .5 - .4898 = .0102$$

and

$$2P = .0204$$

A variation in standard deviations as great as that observed would occur as a chance event about twice in 100 pairs of samples. The receiving clerk is justified in assuming that there is some exterior cause for so great a difference in uniformity of the two car loads of bars. The next step presumably would be a more thorough inspection to discover whether the first car load was more uniform than the specifications of the order required, or whether the second car load was less uniform than required.

Example VIII (Difference of Per Cents of Occurrence).—Several years ago the United States Public Health Service made two separate studies in Buffalo, New York. The first, known as the “public health survey,” was based on information obtained from 22,684 families and the second, known as the “communicable diseases survey” was based on information obtained from 11,709 families. The areas canvassed in the second study were adjacent to those of the first study. Great care was exercised to secure the same representativeness in both samples; hence the two are to be considered as two independent samples from the same universe. The office control cards of the two surveys provided the information needed for a study of vacant dwelling-places in the city. The tabulation of these cards gave the following summary information:

	NO OF DWELLINGS VISITED	NO OF DWELLINGS VACANT	PERCENTAGE VACANT
Public health survey	23,243	559	2.4
Communicable diseases survey .	12,092	383	3.2

Is there a significant difference between the vacancy percentages shown by the two samples?

The standard error of a difference of per cents of occurrence is, from formula 10, page 799,

$$\sigma_{p_1 - p_2} = \sqrt{\frac{p_1 q_1}{N_1} + \frac{p_2 q_2}{N_2}} = \sqrt{\frac{.024 \times .976}{23243} + \frac{.032 \times .968}{12092}} = .0019$$

$$\frac{\text{observed difference}}{\sigma_{p_1 - p_2}} = \frac{p_1 - p_2}{\sigma_{p_1 - p_2}} = \frac{.008}{.0019} = \pm 4.2$$

Hence either P or $2P$ is so small that there is a negligible chance that the difference in the vacancy ratio between the two samples is due to chance variability of sampling alone.

It is not easy to suggest the possible cause of the significant difference in the vacancy ratios of the two samples, because they were so carefully chosen as to be equally representative. The first was taken in February and the second in April of the same year, so that the change in general conditions affecting vacancy must have been negligible. The same agents worked on both studies, and it is possible that in the later study they were less strictly supervised on minor details. The matter of vacancy was not of major importance from the point of view of the health surveys, so that in the second study there may have been a tendency to report as "vacant" some of the addresses from which interviews could not be readily obtained.

Reliability of the Coefficient of Correlation

The meaning and interpretation of the Pearsonian coefficient of correlation have been discussed in considerable detail in chapter XXVII. We are interested here in coefficients computed from samples because a question arises as to the extent of unreliability that can be expected when the relation of the variables in the universe is to be inferred from that found in the sample.

The Standard Error of "r".—The development of a criterion for judging when a value of " r " is significant is less simple than the similar development for the basic measures of a frequency distribution, because the sampling distribution of coefficients of correlation is not normal in shape even though the parent bivariate surface is normal. However, by a process far beyond the level of the present treatise a measure of the standard error of r has been developed:

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}} \text{ (formula 6)}$$

It is ordinarily used like any of the measures of standard error previously explained. Such usage is not fully justified because of the non-normal character of the sampling distribution of r when r approaches unity. Nevertheless, since no alternative exists at the elementary level, a range of $\pm 3 \sigma_r$ on either side of the value of r obtained from a sample is usually taken as the limits within which the r of the universe will fall.

The "z" Transformation.—An alternative method⁹ of measuring the significance of r involves the transformation of r into a new variable called "z." This transformation is accomplished by the equation,

$$z = 1.1513[\log(1 + r) - \log(1 - r)]$$

which can be computed readily with the aid of the table of logarithms in Appendix C.

The effect of this transformation is to change the skewed sampling distribution of r to the normal sampling distribution of z . Variability of z , therefore, can be measured in terms of the standard error of z , since the sampling distribution of z is approximately normal. The probability of a given variation of z can be determined from the Table of Areas. A further advantage of this transformation is the simple form of the standard error of z , i.e.,

$$\sigma_z = \frac{1}{\sqrt{N-3}}$$

The use of the z function as a measure of the significance of r can be understood from an example.

Example IX (Coefficient of Correlation).—Investigation of a sample of 192 undergraduate students at the University of Buffalo showed that there was a positive correlation of .626 between year in school (freshman, sophomore, junior, or senior) and time spent in listening to radio news broadcasts. The question raised is, What can be stated concerning the entire student body with reference to progressive changes in their habits of listening to news broadcasts?

According to formula 6, $\sigma_r = \frac{1-r^2}{\sqrt{N}}$, the standard error of this coefficient is $\pm .044$. A maximum variation of $3\sigma_r$ is $\pm .132$, so the correlation between year in school and radio listening habits with respect to news should not be greater than $+.758$ nor less than $+.494$. This is a rather broad range of variation, but no more exact information is available from the given sample.

This coefficient can be tested by the z function to determine whether the correlation derived from the sample could be obtained by mere chance from a universe in which the correlation was really zero.

⁹ This transformation of r was proposed by R. A. Fisher in *Statistical Methods for Research Workers* (Edinburgh, Scotland: Oliver and Boyd, 1930), pp. 163-71.

The correlation coefficient ($r = .626$) is translated into z for the sake of greater exactitude in the test. The value of z thus computed is

$$z = 1.1513 (\log 1.626 - \log .374) = .735$$

The standard error of z is, $\frac{1}{\sqrt{189}} = .0727$

Then the difference between the z value (.735) and zero is .735 and this difference is more than 10 standard errors

$$\frac{\text{observed difference}}{\sigma_z} = \frac{(.735 - 0)}{.0727} = 10 +$$

and the probability is infinitesimal that the hypothesis of no correlation (the null-hypothesis) is the correct one.

Let us investigate a further question: What is the probability that the correlation in the universe is $+ .758$, the upper limit obtained by the first method? Then

$$z = 1.1513 (\log 1.758 - \log .242) = .991$$

Then

$$\frac{\text{assumed difference}}{\sigma_z} = \frac{.991 - .735}{.0727} = 3.52$$

and

$$P = .5 - .499784 = .000216$$

Thus the chances are about 1 out of 5,000 that a coefficient of .758 would occur as a chance variation from the observed .626, although the first method, the σ_r , indicated the occurrence of .758 as a chance variation once out of 740 trials. The result of the z test is superior because, as previously stated, the sampling distribution of r is not normal and has a definite tendency toward negative skewness as r approaches unity.

SMALL SAMPLES

Differences from Large Samples

Many cases arise in which it is not feasible to obtain samples of the size implied in the preceding formulas. In such circumstances the standard-error formulas are not applicable for two reasons: (1) the sampling distribution of a measure (arithmetic average or standard deviation for example) in a large number of small samples may not

be in the normal form, and (2) the assumption that the value of a measure in the sample is the most probable value of that measure in the universe is not warranted. Modified methods are necessary, therefore, in dealing with small samples. But the distinction between a large sample and a small sample cannot be sharply drawn; hence a question remains as to the occasions for applying the new techniques that are explained in succeeding pages and the occasions for applying the formulas of large samples. Fortunately this problem has a very nice solution. The techniques of small sampling are applicable to large sampling without alteration, although the reverse does not hold.

Large sampling techniques depend upon the normal form of the sampling distributions of the various measures that have been explained. These techniques are designed for two major types of analysis: (1) the determination of the most probable values of measures in a universe when no information is available concerning those measures except what can be learned from a single sample, and (2) the application of tests of significance to measures of a sample when some knowledge is available of the values of such measures in the universe, or some external standard has been established, or two independent samples can be compared.

If the values of a measure are computed from a large number of small samples, the sampling distribution of these values is not necessarily normal. Hence no reliable information about the unknown value of a measure in the universe can be inferred from a single small sample. Therefore the first object of large-sample technique has no counterpart in the field of small samples. The second object, the development of tests of significance, becomes the entire purpose of the study of small samples.

Tests of Significance

The tests of significance that were employed earlier with large samples cannot be used with small samples because, as previously stated, the forms of the sampling distributions of measures of the latter are not dependable. To overcome the uncertainty regarding the shape of sampling distributions for small samples, two new distributions are introduced, known as the "*t*" distribution and the "*z*" distribution. Both of these can be precisely defined by mathematical equations, both are invariant in form like the normal distribution, but neither has the same shape as the normal curve.

The "t" Distribution.—The mathematical background of this distribution was developed in part by an English writer known as "Student," and it is therefore frequently referred to as "Student's distribution." It may be written in the form,¹⁰

$$t = \frac{k_s - k_u}{\sigma} \sqrt{N - m + 1}$$

in which, k_s = the value of any given measure in a sample
 k_u = the value of the same measure in the universe or a standard value of that measure

N = the number of items in the sample

m = the number of predetermined conditions of the sample

$N - m$ = the number of degrees of freedom

$\sigma = \sqrt{\frac{\sum x^2}{N - m}}$ = an estimate of the standard deviation of the universe as computed from the sample

t = a function measuring the value of the difference ($k_s - k_u$) in units of σ and the number of degrees of freedom.

The probability of obtaining any given value of t can be read from Figure 113.¹¹ Find the proper value of $N - m$ on the left scale and the computed value of t on the middle scale. Then the point at which a ruler placed on the proper $N - m$ and t points intersects the "2P" scale at the right will give the probability of obtaining a value of t as great as that observed. If the probability is high, the observed value of t would occur frequently as a chance event and the observed value of k_s could not be said to differ significantly from k_u . On the other hand a value of 2P as low as .05 means that a value of t as great as that observed would occur about once in twenty times as a chance

¹⁰ The t function may be recognized more quickly in the form

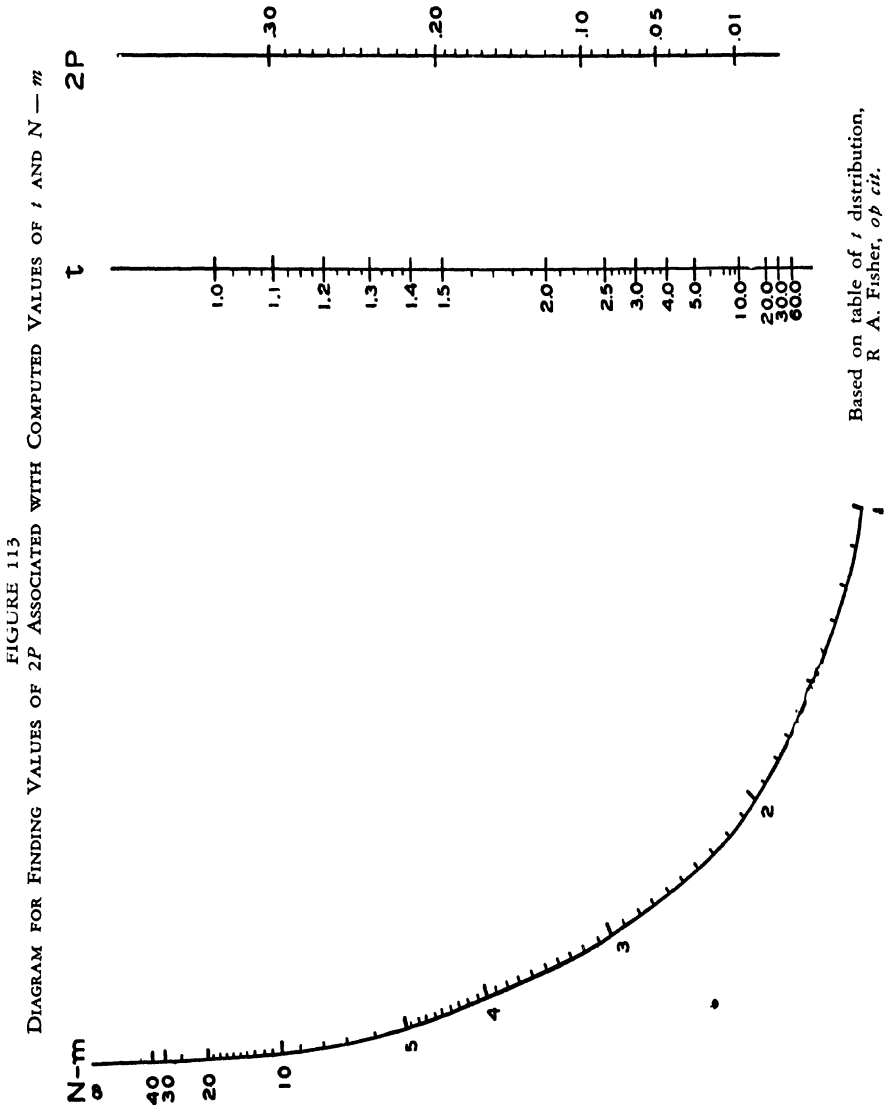
$$t = \frac{k_s - k_u}{\frac{\sigma}{\sqrt{N - m + 1}}}$$

When $m = 1$, as will be the case in Examples X, XI and XII which follow, the denominator term becomes $\frac{\sigma}{\sqrt{N}}$, the standard error of the arithmetic average, and t has the same value as $\frac{x}{\sigma}$ in the normal curve. But the similarity ends at that point. The probabilities of various values of t depend upon the number of degrees of freedom ($N - m$) present in the computation; hence they must be obtained from specially prepared tables or a diagram such as Figure 113.

¹¹ Extended tables of the probability of computed values of t can be found in Karl Pearson, *Tables for Biometricians and Statisticians* (London: Biometrika Office, University College), and tables of the values of t corresponding to various values of P can be found in Fisher, *op. cit.* The essential information of these tables can be obtained from Figure 113, and the form is more convenient.

event. The value of k_s would usually be considered to differ significantly from k_u at the $2P = .05$ level, and values of t at the $2P = .01$ or $2P = .001$ levels are taken as fairly clear evidence of a significant difference between k_s and k_u .

A word of caution is necessary at this point. Figure 113 is constructed to correspond to the tables of R. A. Fisher. That is, the values of $2P$ on the scale at the right of the figure are the probabilities



of obtaining a value as great as $+t$ or as small as $-t$. The use of $2P$ is therefore consistent with the usage followed in determining P and $2P$ from the Table of Areas of the normal curve. If $+t$ alone or $-t$ alone is important in a particular application, the value of P should replace $2P$ in determining the significance of the result.

The t distribution is used in this book to test the significance of a difference of an observed mean from a standard or to test the significance of the difference between two observed means. Therefore k_s becomes M_s (the mean of the sample), k_u becomes M_u (the mean of the universe or the value of a standard) in the first case, and k_s and k_u become M_{s_1} and M_{s_2} in the second case. Some examples of testing means of small samples will demonstrate the computation of t and the use of Figure 113.

Example X.—Eight pieces from an order of 20-gauge uncoated steel sheets were tested for compliance with weight specifications. The United States standard weight for 20-gauge sheets is 1.5 pounds per square foot. The results of the test, in pounds per square foot, were,

X	x	x^2	$\sigma^2 = \frac{.002988}{7} = .0004269$
1.460	.024	.000576	
1.478	.006	.000036	
1.458	.026	.000676	
1.500	.016	.000256	
1.518	.034	.001156	
1.480	.004	.000016	
1.480	.004	.000016	
1.500	.016	.000256	
8)11.874		.002988	
1.484 = M_s			

Then,

$$k_s = M_s = 1.484$$
$$k_u = M_u = 1.5$$
$$\sigma = \sqrt{\frac{\sum x^2}{N - m}} = .0207$$
$$N = 8$$
$$m = 1 \text{ (the use of } M_s \text{ reduces by one the degrees of freedom of this estimate)}$$
$$N - m = 7$$
$$N - m + 1 = 8$$

$$t = \frac{1.484 - 1.5}{.0207} \sqrt{8}$$
$$t = \frac{-.016}{.0207} \times 2.828$$
$$t = 2.19$$

From Figure 113 a value of $t = 2.19$ with seven degrees of freedom has a probability ($2P$) of approximately .07. If the purpose of the test is to discover whether the sheets are too light, the probability

$P = .035$ is wanted. Hence the observed value of M_s would occur some 4 times in 100 as a chance event. The value of t is therefore not certainly significant but the test indicates that the sheets have probably been manufactured to less than standard thickness. The consignment should not be accepted without further inspection.

The minus sign in the numerator of t can be disregarded if the direction of the deviation of M_s from M_u is unimportant. In a test such as this one the average of the sample less than the standard is generally the important variation, but for a particular use excess thickness might be objectionable. In that case one would be interested in whether M_s exceeded M_u significantly. Whatever the purpose of a specific test, the sign of $(M_s - M_u)$ is not used algebraically, but merely as a guide in applying the results of t . Suppose for example that an inspector were interested only in whether the sheets in a shipment were as thick as the specification. Suppose further that the average weight of the sheets in his sample exceeded the specified weight. Then there would be no reason for computing t .

Example XI.—Suppose that from an order of carbon steel floor plates specified as $\frac{1}{2}$ of an inch in thickness four plates were inspected and found to measure .497 inches, .506 inches, .462 inches, and .510 inches. Manufacturers' tables show an average permissible excess thickness of not more than 5 per cent from specified gauge.

The specified gauge is .469 inches, and a positive variation of 5 per cent would permit an average thickness of .492 inches. The test of the four plates would be carried out as follows,

$$\begin{array}{ll}
 M_s = \frac{.497 + .506 + .462 + .510}{4} = .494 & \\
 M_u = .469 \times 1.05 = .492 & t = \frac{.494 - .492}{.0219} \sqrt{1} = .18 \\
 N = 4 & \text{and} \\
 m = 1 & 2P > .3 \text{ (The upper limit of Figure 113)} \\
 N - m + 1 = 4 & P > .15 \\
 \sigma = .0219 &
 \end{array}$$

On the basis of this sample there is no reason to suspect that the thickness of the floor plates exceeds the 5 per cent average excess permitted under the specifications.

Example XII.—An instructor gave the same objective test composed of 40 questions to two groups of statistics tutorial students. The results were as follows:

NO. ANSWERED CORRECTLY

1st Group (7 students).....	31	36	29	24	38	32	34
2nd Group (5 students).....	30	27	30	32	28		

Is there a significant difference in the average ability of the two groups, as measured by this test?

The formula for the t distribution must be changed somewhat when used to test a difference between two means. The form is,

$$t = \frac{M_{s_1} - M_{s_2}}{\sqrt{\frac{\sum x_1^2 + \sum x_2^2}{N_1 + N_2 - (m_1 + m_2)}}} \sqrt{\frac{(N_1 - m_1 + 1)(N_2 - m_2 + 1)}{N_1 + N_2 - (m_1 + m_2) + 2}}$$

in which,

M_{s_1} = the mean of the first sample

M_{s_2} = the mean of the second sample

N_1 = the number of items in the first sample

N_2 = the number of items in the second sample

m_1 = the number of predetermined conditions in the first sample

m_2 = the number of predetermined conditions in the second sample

$\sum x_1^2$ = the sum of squares of deviations of items in the first sample from M_{s_1}

$\sum x_2^2$ = the sum of squares of deviations of items in the second sample from M_{s_2}

$N_1 + N_2 - (m_1 + m_2)$ = number of degrees of freedom

When m_1 and m_2 are each equal to unity, as is the case in testing the difference of two averages, the formula becomes,

$$t = \frac{M_{s_1} - M_{s_2}}{\sqrt{\frac{\sum x_1^2 + \sum x_2^2}{N_1 + N_2 - 2}}} \sqrt{\frac{N_1 N_2}{N_1 + N_2}}$$

For the given example then,

$M_{s_1} = 32$

$M_{s_2} = 29.4$

$N_1 = 7$

$N_2 = 5$

$m_1 = 1$ } (M_{s_1} and M_{s_2} are computed
 $m_2 = 1$ } from the samples)

$\sum x_1^2 = 130$

$\sum x_2^2 = 15.2$

$N_1 + N_2 - (m_1 + m_2) = 10$

$$\begin{aligned} t &= \frac{32 - 29.4}{\sqrt{\frac{130 + 15.2}{10}}} \sqrt{\frac{7 \times 5}{12}} \\ &= \frac{2.6}{\sqrt{14.52}} \sqrt{2.9167} \\ &= \frac{2.6 \times 1.71}{3.81} \end{aligned}$$

$t = 1.17$

In Figure 113 for ten degrees of freedom and $t=1.17$, the value of $2P$ is between .20 and .30. That is, a difference of means as great as this might occur as a chance event in 2 or 3 out of every 10 pairs of samples. The difference between the two groups of tutorial students is therefore not significant.

*The "z" Distribution.*¹²—This distribution was developed by R. A. Fisher,¹³ and serves the same purpose for a difference between two observed standard deviations of two small samples that the t distribution serves for a difference between two observed means. The formula is,

$$z = 1.1513[\log \sigma_1^2 - \log \sigma_2^2]$$

or

$$z = 1.1513 \left[\log \frac{\sum x_1^2}{N_1 - m_1} - \log \frac{\sum x_2^2}{N_2 - m_2} \right]$$

in which all of the symbols on the right side of the formula have the same meaning as in the t formula on the preceding page.

The function z is distributed according to an equation that differs somewhat from the normal form and also from the form of the t distribution. It therefore requires separate tables for determining significance levels. Such tables have been replaced in this book by two diagrams, Figures 114 and 115, which give the number of degrees of freedom of the first sample ($N_1 - m_1$) on the left scale, the number of degrees of freedom of the second sample ($N_2 - m_2$) on the middle scale, and the values of z corresponding to the .05 (Fig. 114) and .01 (Fig. 115) levels of significance on the scale at the right. The nature of the distribution of z is such that a unified scale is obtained in each case by allowing the degrees-of-freedom scale of the second sample to cross the z scale as shown in the diagrams. Either scale is used in the following manner: having found the points corresponding to the proper degrees of freedom on the left and middle scales, a ruler is laid across the two points so that it intersects the z scale. This intersection gives the value of z that would be expected as a chance event once in 20 trials in Figure 114 and once in 100 trials in Figure 115.

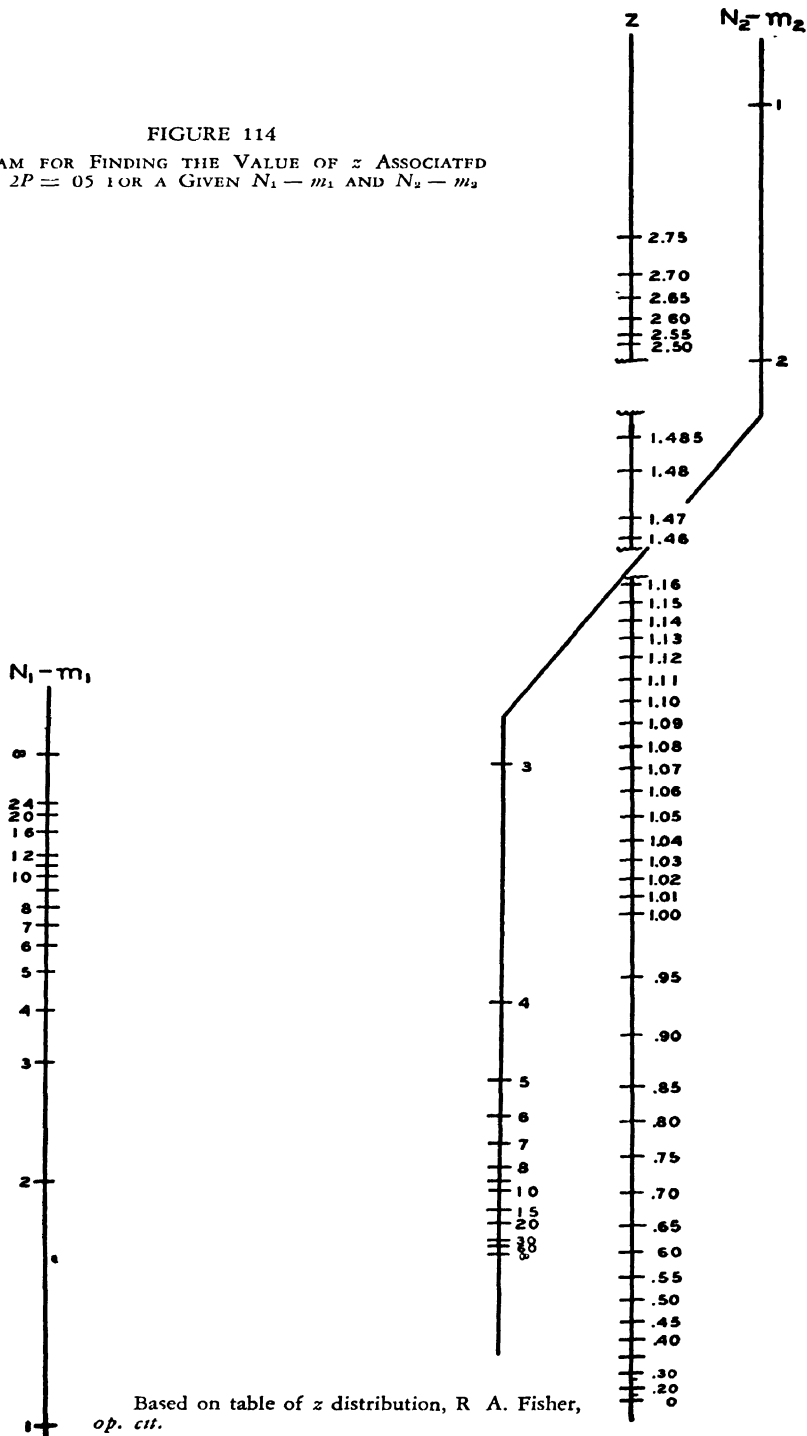
In all cases in which Figures 114 and 115 are used the ($N_1 - m_1$) scale should go with the sample having the larger standard deviation.

¹² The z distribution used in small-sample tests is to be distinguished from the z transformation used in measuring the significance of a coefficient of correlation.

¹³ *Op. cit.*, pp. 194-215.

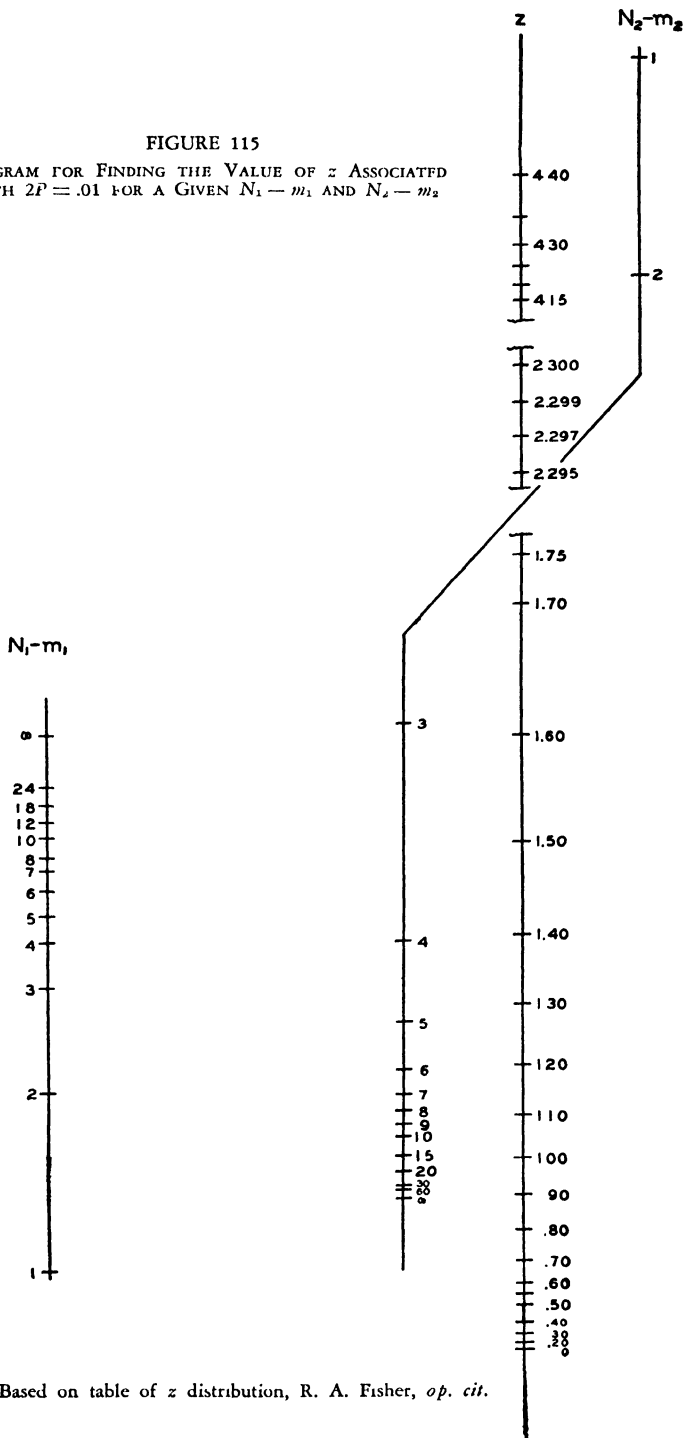
FIGURE 114

DIAGRAM FOR FINDING THE VALUE OF z ASSOCIATED
WITH $2P = .05$ FOR A GIVEN $N_1 - m_1$ AND $N_2 - m_2$



Based on table of z distribution, R. A. Fisher,
op. cit.

FIGURE 115
 DIAGRAM FOR FINDING THE VALUE OF z ASSOCIATED
 WITH $2P = .01$ FOR A GIVEN $N_1 - m_1$ AND $N_2 - m_2$



Based on table of z distribution, R. A. Fisher, *op. cit.*

Example XIII.—Suppose the question were raised concerning the students in the two tutorial groups of Example XII, Is there more uniformity of ability in the second group than in the first as measured by the test results?

From the data given in Example XII it is clear that the sample of seven students has the larger standard deviation. Therefore,

$$\sigma_1^2 = \frac{130}{6} = 21.67; \quad \sigma_2^2 = \frac{15.2}{4} = 3.8$$

and

$$\begin{aligned} z &= 1.1513[\log 21.67 - \log 3.8] \\ z &= .87 \end{aligned}$$

When $(N_1 - m_1) = 6$ and $(N_2 - m_2) = 4$, the value of z at the 5 per cent level in Figure 114 is .92, and at the 1 per cent level in Figure 115 is 1.37. Therefore the computed value of z could be expected to occur as a chance event more frequently than once in 20 trials. The hypothesis that there is a significant difference in the uniformity of mental development of the two groups of students is not demonstrated, according to the test by the z distribution.

Example XIV.—Suppose that two separate samples of a consignment of 14-ounce copper sheeting yielded the following results:

THICKNESS IN INCHES	NO. OF PIECES MEASURED	
	Sample 1	Sample 2
.0160-.0169	1	0
.0170-.0179	2	1
.0180-.0189	10	12
.0190-.0199	8	6
.0200-.0209	4	1
	25	20

The United States Bureau of Standards places the theoretical thickness of 14-ounce sheets at .0189 inches. The means of both samples are fully up to this specification, but the question of uniformity is to be determined by the z test. The required data for the z formula are,

$$\begin{aligned} \sigma_1^2 &= .00000101 & \sigma_2^2 &= .00000045 \\ z &= 1.1513[\log .00000101 - \log .00000045] \\ z &= .40 \end{aligned}$$

From Figures 114 and 115 for $(N_1 - m_1) = 24$ and $(N_2 - m_2) = 19$, if $2P = .05$, $z = .37$; if $2P = .01$, $z = .54$. The difference in uniformity of

the two samples is therefore not conclusively shown, although it is strongly suggested.

Before leaving this example another question should be raised, namely, Could these two samples be tested by large-sample measures? The formula for the standard error of the difference of two standard deviations is formula 9 (page 799),

$$\begin{aligned}\sigma_{\sigma_1 - \sigma_2} &= \sqrt{\frac{\sigma_1^2}{2N_1} + \frac{\sigma_2^2}{2N_2}} \\ &= \sqrt{\frac{.00000101}{50} + \frac{.00000045}{40}} \\ &= .000177\end{aligned}$$

Then, for use in the Table of Areas,

$$\frac{x}{\sigma} = \frac{\sigma_1 - \sigma_2}{\sigma_{\sigma_1 - \sigma_2}} = \frac{.001005 - .000671}{.000177} = \pm 1.89$$

If $\frac{x}{\sigma} = 1.89$, $P = .5 - .4706 = .0294$, and $2P = .0588$. That is, the difference between these two standard deviations would occur as a chance event a little more than once in 20 trials. This is fairly close to the result obtained by the z test.

The closeness of these two results offers some evidence concerning the dividing line between large samples and small samples. In this case samples of 20 and 25 items are so related that the sampling distribution of the differences of standard deviations can be assumed to be normal without distortion of results. One must not, however, generalize that all samples of 20 or more are large samples. The satisfactory result obtained in this example depends to a large extent upon the uniform character of the material being sampled and the consequent small values of σ_1 and σ_2 .

One general statement can be based on the results of this example. The number of cases required for the use of large-sampling technique will depend upon the homogeneous character of the universe involved. Specifically, in the field of industrial inspection and testing, samples with as few as 20 items can be analyzed by large sampling methods.

VARIANCE ANALYSIS

In recent years statisticians have devoted a great deal of attention to a new subject called variance analysis. *The variance of a distribution*

is simply the square of the standard deviation. Variance analysis then is a study in terms of σ^2 . The new feature consists in the separation of the total dispersion of a distribution into two categories. The partial dispersions are then compared to determine whether the two differ significantly or within the limits of events due to chance.

Variance analysis can be used wherever two separately measurable independent causes of dispersion affect a single set of data. The method has its greatest usefulness in the field of agricultural experimentation, but can be applied to a limited range of business problems. Suppose that a farmer wished to determine which of six varieties of seed corn would produce the best yield on his land. He might prepare 30 acres identically, plant each of six 5-acre plots with a different variety of seed, cultivate all plots similarly, and note the yield of each variety. The results would not be conclusive, however, because soil differences between plots might be more important than differences between varieties of seed. This method combines two causes of dispersion or variance: (1) the variance attributable to the difference in the varieties of corn and (2) the variance due to differences in the fertility of the 5-acre plots of land. The first variance is of interest in the experiment; the second is to be eliminated if possible.

This could be accomplished by planting a large number of such 30-acre plots. The variance in soil fertility would affect each variety of seed corn at random. The yields of the several varieties from the large number of plots would form samples affected equally by variance due to differences in soil fertility. By the usual methods the averages of the resulting six samples could be tested for significant differences in yield of the six varieties of seed.

Since this method can seldom be carried out literally as described above, a small-scale equivalent of random sampling is substituted. The farmer might divide his plot into 30 one-acre sections and then plant each variety of seed in 5 random non-contiguous acres of his plot. Differences due to soil fertility would thereby be reduced to a minimum and the variations in average yield would be caused mainly by differences in the varieties of seed corn. But the variance between the varieties must be compared with the variance within the individual varieties in order to determine whether the former is significant.

Before proceeding with the methods of measuring variance and the tests of significance, another illustration will be introduced to show some of the limitations upon this type of analysis.

Suppose that a class of 125 students in a university were divided into five sections, each taught by a different instructor. If the same examination were given to the five sections, and all papers were graded by some impartial person, the differences in average grades between sections would indicate differences in the average ability of the sections combined with differences in the effectiveness of the instructors. Tests of significance applied to the average grades of the sections could not be interpreted with reference to either cause. Now suppose that the teachers were shifted at the end of each month in such a way that at the end of five months each section would have been under the tutelage of each instructor for one month. The variance due to effectiveness of instructors would thereby be reduced to a minimum and the main cause of variance in average grades between sections would be difference in ability of the sections. The significance of the variance between sections could be tested by comparing it with the variance within the sections.

While it may appear that this example is an exact counterpart of the varieties of seed corn, there are in fact several reasons why the second example is not well adapted to variance analysis. The teachers may not have been equally attentive to duty during all of the five months. The students may not have given the same attention to this course during all of the five months. The teachers may have had variable familiarity with different parts of the course. The grading of the examination papers may not have been uniform as between sections. All such factors were to be eliminated as a cause of variance by the device of alternating teachers of the sections. But human material cannot be regularized in the same way as acres of land. Hence the full conditions for variance analysis are not met and for that reason the results of an experiment such as the revolving of teachers are of doubtful value regardless of the findings of a statistical test of significance. It was with due regard to these circumstances that the statement was made at the beginning that variance analysis has its greatest usefulness in the field of agricultural experimentation.

Nevertheless in the field of business statistics some situations arise that call for the use of variance analysis. Accordingly the method will be explained by means of an example.

Table 163 contains a list of percentages of advertising*expense to sales of 89 concerns separated into six groups according to size as measured by sales. The data are the equivalent of a segment of a

national investigation made by an advertising association and pertain to advertising budgets for the year 1937. The 89 concerns were all engaged in manufacturing but no separation was made according to type of product. Hence each of the six groups contains a random set of concerns in different fields of manufacturing and from various sections of the country. To make the groups truly random a much larger sample would be necessary. The smaller sample is used here for illustrative purposes to reduce computation, and to set up a border-line case between large sampling and small sampling. The analysis of the variance will be the same regardless of the size of the sample except for the question of degrees of freedom which will be discussed later, but the test of significance will be different. Both methods will be shown for the example.

The question to be raised concerning these percentages is, Are the observed differences in the ratios of advertising cost to sales for concerns of different size significant, or are they merely a chance accompaniment of the variations of the ratios within the six size groups? In other words, do the percentages spent for advertising decrease significantly with increasing sales? The variance within the six size groups is computed as follows:

1. For each column of ratios, take the sum of the squares of the deviations from the average of that column. (These sums are shown for each column and should be verified by the student.)

2. The total of the six sums is divided by the number of ratios in the table to obtain the variance within the several groups. The result is 1.8 per cent as shown at the lower left of the table.

This is the average variation or variance due to differences of percentages spent for advertising by concerns in the same size group. Size of concern is eliminated as a cause of this variance.

The next task is to measure the variance due to difference in size of concern. The steps are:

1. Take the difference of each group average from the grand average of the entire table (shown at the lower right of the table).

2. Square these deviations.

3. Weight these squared deviations by the number of ratios in each group to obtain the total amount of deviation due to each group.

4. Sum the weighted squared deviations and divide by the *number of groups* to obtain the total variance between the averages of the groups.

TABLE 163

COMPUTATION OF VARIANCE OF THE RATIO OF ADVERTISING EXPENSE TO SALES IN PER CENTS, FOR 89 CONCERNS IN SIX SIZE GROUPS,
Size Based on Sales (in million dollars)

LESS THAN .5	.5-1.0	1.0-2.0	2.0-5.0	5.0-10.0	10.0 AND OVER
3.0	2.3	1.1	8	3.7	6.8
5.1	3.3	2.7	3.0	1.1	2.3
1.9	1.8	2.3	1.9	2.3	4
6.6	1.0	4.6	2.1	1.4	.7
2.8	4.3	6.3	.7	1.8	1.6
3.3	5.7	3.0	1.6	3.8	1.6
3.1	3.0	2.1	1.1	.7	1.8
2.4	2.2	1.0	.4	1.6	3.0
1.1	2.1	1.7	2.7	1.2	.9
8	3.4	2.0	1.2	4.5	1.1
3.2	1.4	2.7	3.8	2.0	1.2
5.0	1.6	.7	.9	1.1	.7
4.4	2.0	1.3	1.4	1.9	3.6
6.8	2.3	2.6	1.8	...	2.4
4.0	3.2	...	2.9	...	...
2.1		...	1.6		
2.3					
Average ratio 3.4	2.6	2.4	1.7	2.1	2.0
Sum of squared deviations from av. 47.47	20.74	29.13	13.61	16.90	35.81

VARIANCE BETWEEN THE GROUPS

VARIANCE WITHIN THE GROUPS	DEVIATIONS OF GROUP AVERAGES FROM GRAND AVERAGE (2.4)	DEVIATION SQUARED	NO. OF RATIOS IN GROUP	WEIGHTED SQUARED DEVIATIONS
47.47	1.0	1.00	17	17.00
20.74	.2	.04	15	60
29.13	0	...	14	...
13.61	-.7	.49	16	7.84
16.90	-3	.09	13	1.17
35.81	-.4	.16	14	2.24
89) 163.66				6) 28.85
18				4.8

This is the variance due to difference in the size of concerns. By using the average of each group the variance within the groups has been eliminated. If the variance between the groups is no greater than the variance within the groups, one can safely assume that the same chance variability that produced the variance within the groups would explain the variance between the groups. Also, if the variance between the groups were less than the variance within the groups, one would conclude that some interrelationship between the groups had operated

to reduce the effect of the chance variability present within the groups. On the other hand, if the variance between the groups exceeds by a significant amount the variance within the groups, some causal factor not present within the groups must be operative. The variance within the groups serves as a yardstick to determine whether some causal factor is present in the variance between the groups.

In this example the variance within the groups is 1.8 per cent and the variance between the groups is 4.8 per cent. Apparently then some external cause is present. This conclusion must be tested, however, before it can be accepted finally. That is, the two measures of variance must be tested to determine whether the difference is significant. The conclusion will be tested first on the assumption that we are dealing with a large sample and then on the assumption that we are dealing with a small sample.

In large-sample theory the question is simply whether the two variances differ significantly. The formula for the standard error of the difference of two standard deviations is,

$$\sigma_{\sigma_1 - \sigma_2} = \sqrt{\frac{\sigma_1^2}{2N} + \frac{\sigma_2^2}{2N}}$$

then
$$\sigma_{\sigma_1 - \sigma_2} = \sqrt{\frac{4.8}{2 \times 6} + \frac{1.8}{2 \times 89}} = .64$$

and
$$\frac{x}{\sigma} = \frac{\sigma_1 - \sigma_2}{\sigma_{\sigma_1 - \sigma_2}} = \frac{2.20 - 1.36}{.64} = \frac{.84}{.64} = \pm 1.3$$

When $\frac{x}{\sigma} = 1.3$, $P = .5 - .4032 = .0968$ (Table of Areas of the normal curve) and $2P = .1936$. Hence the difference could occur as a chance event in about 1 out of every 5 cases, and is not significant.

According to this test the conclusion that percentage of sales expended for advertising varies with size of industrial concern is unwarranted on the evidence at hand. The assumption that the methods of large sampling can be applied to the problem may be faulty. With not more than 17 concerns represented in any group and only six groups, there is no certainty that the sampling distribution of the difference between the two standard deviations is normal.

The safer procedure, therefore, is to shift to small sample techniques. Degrees of freedom must be substituted for number of items in computing the variances in Table 163. Then the variance within the

groups is $\frac{163.66}{83} = 2.0$. Six degrees of freedom are lost because the average of each group is computed from the data in the group. One degree of freedom is lost in computing the grand average; hence the variance between groups is $\frac{28.85}{5} = 5.8$.

To test the significance of the differences between these two variances, the z distribution must be employed. The formula for z is,

$$z = 1.1513[\log \sigma_1^2 - \log \sigma_2^2]$$

in which σ_1 is always the greater variance. Then,

$$z = 1.1513[\log 5.8 - \log 2.0] = .53$$

The probability of obtaining a value of $z = .53$ with $(N_1 - m_1) = 5$ and $(N_2 - m_2) = 83$ can be determined from Figures 114 or 115. From Figure 114 a value of z of about .43 would occur once in 20 trials as a chance event. From Figure 115 a value of z of about .57 would occur once in 100 trials as a chance event. The computed value, $z = .53$, therefore would occur a little more frequently than the latter, or the probability of the observed value of z occurring as a chance event is between $P = .05$ and $P = .01$.

This test indicates that the variance between groups is probably significant. Thus by the use of small sampling methods much more rigor is introduced and the uncertainties as to the normality of the sampling distribution of differences of standard deviations is avoided. In all likelihood investigation of a larger sample would confirm the conclusion that the difference between the two variances is indicative of a significant relation between size of concern and percentage of sales expended for advertising.

PROBLEMS

1. Explain (a) the meaning of "sampling distribution," (b) the difference between "standard deviation" and "standard error."
2. Why does the development in the early part of this chapter depend upon the normality of sampling distributions?
3. The cost of building an identical house in each of 77 cities in various parts of the United States at the close of 1940 appears in the 1941 Statistical Supplement of the *Federal Home Loan Bank Review*. The following is a summary of the data.

Average cost	\$6,029
Standard deviation of cost	\$ 459
Number of cities.....	77

- a) According to this sample what is the range within which the chances are 2 to 1 that the actual average cost of building this urban dwelling will fall? The range within which the chances are 19 to 20? The range within which the chances are 369 to 370? The range within which the chances are 99 to 100?
- b) What relation between sample and universe is implied in the answers to (a)?
4. Compute Q_1 and Q_3 for the distribution of age of automobiles in Problem 8, pages 782-83. What are the upper and lower limits of the interquartile range of the universe represented by this sample?
5. What is the probability that the cost of building the house described in Problem 3 was as low as \$6,000 at the end of 1940?
6. A student survey (made in 1938) of 4,403 automobiles in parking lots in Buffalo showed that 33.3 per cent of the cars were Fords. Data obtained subsequently from official sources showed that 29.7 per cent of the cars registered in Buffalo were Fords. Was the sample representative with respect to inclusion of Ford cars?
7. Tabulation of data parallel to those described in Problem 3 at the close of the first quarter of 1941 gave the following,

Average cost of house	\$6,232
Standard deviation of cost	\$ 504
Number of cities.....	68

Is this a significant increase in cost of constructing a house or may it be merely a sampling variation?

8. A survey similar to that described in Problem 6 was made in 1939. (Exactly one year elapsed between the two surveys.) The 1939 survey included 5,041 automobiles and 28.8 per cent were Fords. Was the difference between the two samples significant? If so, what are some likely causes of the difference? In your answer take into account the additional information available in Problem 6.
9. Results for a long period of time show that the operator of a bolt-threading machine will spoil 50 bolts per 1,000. A new operator was put on the machine and his first run of 1,000 bolts showed 100 spoiled. Would you conclude that the new operator was unsatisfactory?
10. Two department stores in the same city carry about the same line of goods. In one department the following results were obtained for one month.

	STORE A	STORE B
Average sales check	\$4 236	\$4.102
No. of checks	1,200	800
Standard deviation	\$1.36	\$1.56

- a) Assuming that size of sales check indicates difference in economic status of customers, approximately what are the chances that these two departments draw their customers from the same economic levels? That is, were the samples drawn from the same universe?
 - b) Supposing that the two samples represented the same store in two successive years, what additional complication in inferring the conditions of the universe from the samples would be introduced thereby?
11. Investigate the probability that the coefficient of correlation between rents and incomes of families in Columbus, Ohio, is as low as .85 (see Table 150, chapter XXVII, p. 726).
 12. Investigate the probability that the coefficient of correlation between prices of common stocks and earnings per share of all chemical manufacturers is as low as .90 (see chapter XXVII, p. 719). Use the z test.
 13. The average sales per employee-hour in 12 retail variety chains in 1937 was \$2.18. Similar results from the same chains in 1938 were as follows:

SALES PER EMPLOYEE-HOUR IN 1938

\$2.76	1 74	2 05
1 74	2 13	2 05
2 01	2 17	2 97
2.09	2 63	2 66

Source: See chapter XXVII, Problem 2, page 746.

Did average sales per employee-hour in retail variety chains differ significantly in 1938 from the 1937 results?

14. An investigation of 17 department and specialty stores in the same community produced the following data:

SIZE OF STORE (Sales)	No. OF STORES	AVERAGE ANNUAL SALES PER SALES PERSON	STANDARD DEVIATION OF SALES PER SALES PERSON
Less than \$500,000	10	\$4,422	\$750
\$500,000 and over	7	5,128	293

and can one conclude from these data that there is a significant difference between sales per sales person in large and small stores?

15. A sample of farm customers of privately owned and municipally owned electric lighting companies in New York State showed the following concerning number of kilowatt hours of current consumed per month.

	PRIVATELY OWNED COMPANIES	MUNICIPALLY OWNED COMPANIES
No. of customers	73	33
Average consumption	38 6 k w h.	36 7 k.w.h.
Standard deviation	11.3 k.w.h.	9 6 k.w.h.

The conclusion of the study was that the customers of municipally owned companies consume current in more uniform quantities than the customers of privately owned companies. Is this conclusion justified?

16. Four sets each consisting of five different advertisements were prepared so that all words identifying the advertiser directly were removed. Each set contained advertisements of (a) cigarettes, (b) liquor, (c) food, (d) household equipment, and (e) automotive products. The sets were shown to an equal number of men and women and to different age groups. Each person saw one set of advertisements. The results from 52 interviews are recorded below.

NO. OF ADVERTISEMENTS RECOGNIZED			
SET I	SET II	SET III	SET IV
3	2	1	3
4	4	2	3
2	3	2	4
3	1	4	2
1	3	0	0
3	4	3	3
5	2	2	4
4	5	1	3
3	3	3	5
0	2	0	2
2	3	2	4
3	0	4	5
	1	0	3
	2		

A conscious effort was made to select the advertisements so that the results of the four sets would be comparable. Yet a question arose in the analysis concerning the comparability of the results.

- a) Test the variability between the sets in terms of the variability within the sets.
- b) Does this problem meet the requirements for variance analysis? Discuss.

REFERENCES

BROWN, THEODORE H., "The Use of Statistical Techniques in Certain Problems of Market Research," *Business Research Studies No. 12*. Boston: Harvard University Graduate School of Business Administration. Vol. XXII (May, 1935).

DAVIES, GEORGE R., and YODER, DALE, *Business Statistics*. New York: John Wiley and Sons, 1937.

Page 501 contains a graph of the "t" distribution in a form different from that presented in this text.

DUNLAP, J. W., and KURTZ, A. K., *Handbook of Statistical Nomographs, Tables and Formulas*. New York: World Book Co., 1932.

Contains nomographs for determining the values of standard errors of basic measures of frequency distributions and coefficients of correlation.

FISHER, R. A., *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1930.

The entire text is indispensable for advanced study of the principles of sampling. Complete tables of the "*t*" distribution, the "*z*" distribution and the "*z*" transformation are included.

FISHER, R. A., and YATES, F., *Statistical Tables for Biological, Agricultural, and Medical Statistics*. Edinburgh: Oliver and Boyd, 1938.

KENNEY, JOHN F., *Mathematics of Statistics*. New York: D. Van Nostrand Co., Inc., 1939.

Chapters VI and VII of Part II give a mathematical presentation of the theory and formulas of sampling.

MILLS, FREDRICK C., *Statistical Methods*. New York: Henry Holt and Co., 1938.

Variance analysis is applied to a wide range of problems, pages 490–529.

SHEWHART, W. A., *Economic Control of Quality of Manufactured Product*. New York: D. Van Nostrand Co., Inc., 1931.

The entire content deals with the principles of sampling and their application in the control of various manufacturing processes.

YULE, G. UDNY, and KENDALL, M. G., *An Introduction to the Theory of Statistics*. London: Charles Griffin and Co., Ltd., 1937.

Pages 332–461 give a complete presentation of the theory of sampling and a demonstration of the use of various sampling measures.

CHAPTER XXX

PRESENTATION OF THE RESULTS OF STATISTICAL INVESTIGATION

INTRODUCTION

THE steps in statistical investigation as set forth in chapter III are collection, tabulation, analysis, and presentation. The discussion of the first three of these is the content of the preceding chapters of the book. This chapter deals solely with the last stage of the investigation procedure. Only in the case of an individual carrying on research for his personal satisfaction with no intention of making his results known would this final step be omitted. Under all other circumstances the results of investigation would be reported in written form; hence the discussion which follows is concerned with the method of preparing written reports.

Importance of Presentation

Too little attention is often given to the preparation of reports, with the result that valuable work may be so poorly presented that it receives slight recognition. This is true of reports of external investigation presented for general use as well as of those prepared within business concerns. The latter type of report is particularly important because it usually relates to a research undertaken to throw light on some problem of management, and a poor presentation of the results of the work may cause management to discard valuable information or misinterpret the facts. When this happens statistical research which should be a valuable adjunct of the managerial function becomes worse than useless. On the other hand, poor presentation of external research may sometimes be overcome in the long run, if the report includes complete and accurate data that can be used later by other investigators who may present and interpret them more effectively.

Scope of Present Discussion

Written reports range in length from ordinary letters to full-sized books. In this chapter the purpose is to place no emphasis on report

writing at either of these extremes, but rather on the type of reports made in business. Training in the preparation of such reports can be made an adjunct to classwork for business students.

Business report writing has two major facets: (1) applied English, (2) the explanation of statistical techniques and interpretation of results. The second point only will be emphasized in this chapter, except for some general references to style and accuracy. No attempt will be made to cover the broad subject of thesis writing. It is assumed that ability to use the English language effectively is an indispensable basis for the discussion of the parts of a business report as well as of a report in any other field of knowledge.

THE WRITER—READER RELATION

Research is never routine in character. The very name implies that the investigator is delving into something new, that he is at the very least testing known facts in new combinations. It follows that the presentation of the results of research is a perpetual challenge to the author to tell his story effectively. The author of a report is usually solely responsible for the success of his undertaking and success is measured by neither hours spent nor pages produced but by the feeling of a task well done. A task well done is probably in the last analysis one which completely satisfies the scientific judgment of the author.

Scientific Attitude

A dispassionate attitude of scientific investigation is a prerequisite for the presentation of any research work. The author must be thorough, inquisitive, unbiased, self-critical, and skeptical. These qualities are by no means the natural endowment of every individual; they must be developed by painstaking mental discipline. Experience in research work and report writing is the only method of making these qualities an automatic part of one's mental equipment, but to begin training the mind properly from the start is as essential here as is proper stance and rhythm to a person learning to play golf.

The writer of a research report should never give way to the natural tendency toward enthusiasm. If he has been particularly successful in obtaining new knowledge, it is only natural that the importance of his work will be magnified in his own thinking. This tendency must be suppressed in writing a report of the work. It should be clear also

that there is no place for special pleading in research reports. The author must be certain that the conclusions have been deduced from the facts and that the facts have not been distorted to fit any preconceived opinions masquerading as conclusions.

The Reader

One of the most difficult phases of report writing is to determine the exact level at which the writing should proceed. The tendency of the report writer, steeped as he is in knowledge of all aspects of his subject, is to write learnedly. The primary object is not to appear erudite but to gain an audience. While this could be accomplished readily by presenting conclusions pleasing to the reader, such practice must, of course, be condemned.

The investigator is thoroughly familiar with every detail of the work and from his point of view all of the valuable conclusions could be presented in a paragraph or even in a single sentence. But a report is written for the use of others who are not familiar with the research work which has preceded the writing. For that reason it is always necessary to keep in mind the readers to whom a report will go. The organization, content, and style of a report will depend upon the attitude and knowledge of readers as determined by (1) whether their interest centers in the details of the work or only in the final conclusions and (2) the extent to which a knowledge of statistical methods can be assumed.

An internal report on some phase of selling would have a certain emphasis if directed to the sales manager, but a quite different emphasis if it were to come to the attention of the board of directors. A research agency in preparing a report addressed solely to the firm sponsoring the work would omit much in the way of method and supporting evidence that would have to be included if the report were to be published. For example, a report of the results of testing the effectiveness of advertising a certain product on bill boards, in magazines, or by direct mail might include merely a statement of the findings, if made to the manufacturer of the product. On the other hand a magazine article based on this investigation would include an explanation of the method of testing the three advertising media, a justification of the validity of the sampling method, a statement of the actual conduct of the test, a full presentation of the findings, and some conclusions as to the general applicability of the results.

Distinctions of organization and style must be taken into account in many directions in adjusting a report to its readers. If they possess a statistical background, the description of the process of collection and analysis can be phrased to make use of that knowledge. But the same report addressed to persons unfamiliar with statistical methods would need to include descriptions of all statistical processes. For example, in the latter case a report based upon sample data would necessarily include an explanation of how results representing an entire universe can be obtained from analysis of a small sample drawn from that universe. If the same report were addressed to statisticians, it would be necessary merely to explain the steps taken to insure representativeness.

In the final analysis the task of the report writer is to concentrate on a clear and decisive statement of the facts as he has found them, avoiding either over- or understatement, and taking special pains to keep constantly in mind the extent of knowledge of his subject possessed by the reader or readers to whom his report is addressed.

The only guide to the level at which a report should be written is combined judgment and experience. This is only another way of saying that the beginner must depend upon a process of trial and error while he acquires experience and develops judgment. For this reason it is important that students practice report writing to avoid finding themselves in a defenseless position, when after graduation they are called upon to prepare reports in some business capacity.

REQUIREMENTS OF A REPORT

The remainder of the chapter is confined to a more specific point of view than that which characterized the preceding discussion. We propose to explain in detail the standards of report writing that must be inculcated into the mind of the beginner as a sound basis for the report writing he will be called upon to do in college and as a proper background for any report writing he may do in a business capacity. As previously indicated the latter type of report commonly gives less attention to formal requirements and more attention to the reader point of view. Nevertheless a knowledge of the requirements of a complete report is necessary in order to be able to use good judgment in preparing reports in abridged form when the occasion arises to do so.

The specific practices in the construction of reports call for detailed discussion. These things must be consciously developed at the outset, although at a later stage they become automatic.

Accuracy

Throughout preceding chapters emphasis has been placed upon accuracy in statistical work. That requirement is no less applicable to report writing and in truth needs special emphasis.

In particular, accuracy should be observed in the introduction of quotations and references. A quotation from either printed material or oral authority must be properly set apart in the text and the source indicated. All references must be given in enough detail so that they can be consulted readily by the reader of the report. The practice followed in this book is probably the most convenient method for student reports, i.e., footnotes numbered consecutively, placed at the bottom of the page on which the reference occurs and separated from the text. The approved form for footnote references can be obtained from any style manual.¹

Adequacy

A report may be prepared to answer a specific question or to marshal facts concerning a business or economic problem. In either type of report the work must be completely planned and the execution of the plan must lead to an adequate solution of the problem. The beginner is likely to be satisfied with *any* solution but more experience will demonstrate the necessity of investigating various possible solutions instead of seizing the first one that presents itself.

Investigating the subject adequately is not synonymous with presenting everything that has been learned during the research process. Unsuitable data and methods of analysis must be omitted, leaving a connected logical development in the final presentation.

The attempt to give an adequate presentation sometimes leads to a very tedious report. This results from needless elaboration of the obvious or from permitting extraneous or collateral matter to assume unwarranted importance. The report writer can guard against this by laying down in the beginning a rigid plan of writing and forcing him-

¹ See particularly *A Manual of Style* (10th edition; Chicago: University of Chicago Press, 1937), pp. 125-30.

self to adhere to it until the complete report has been drafted. It is time then to go over the first draft, adding to parts not adequately presented and deleting whatever detracts from the smooth flow of the subject matter.

Soundness

The question of soundness relates to the interpretation of the results of research. Only in the simplest kind of research work will the analysis lead to results which point infallibly to a single conclusion. More commonly the results appear in the form of summary figures which must be related to non-numerical information in order to draw valid conclusions.

Students often mistake a description of the process of analysis for the drawing of conclusions. Interpretation or conclusions should answer questions such as: What new relationship has been discovered? What does the comparison of the figures mean? What is the meaning of the difference in rate of growth of the curves? How can the general relationship be applied to particular situations? The exact form in which questions of this character will be raised will depend upon the type and purpose of different investigations. This process of comparison, elimination, and inference, known as interpretation, causes the greatest difficulty in report writing and gives rise to the greatest amount of statistical misinformation.

This is the point at which the most guidance is needed, but unfortunately little specific assistance can be given. In preceding chapters the use and misuse of various techniques have been explained. Knowledge of the logical background of the discussions of ratios, averages, index numbers, time series, and the tools of more advanced analysis should provide the basis for drawing sound conclusions when any of these techniques have been employed in research work. Every novice is certain to discover, however, that there is a hiatus between his knowledge of the use of statistical techniques in general and his ability to select and apply those techniques discriminatingly to the material collected for the solution of a particular problem. This "missing link" is usually not of statistical origin in the narrow sense, but relates to the lack of general knowledge of the subject under investigation. Statistical results must be interpreted in relation to the particular setting provided by the subject matter in each case. This necessitates a broad knowledge of business affairs as a background and an intensive study

of the factors surrounding the particular problem with which the statistician is dealing.

Reference to the data concerning wages paid to workers engaged in manufacturing explosives, as presented in Table 16-A, page 151 in chapter VIII illustrates this point. It will readily be noted that the total wage distribution is "flat-topped," with modal concentration in two wage classes, 62.5–67.5 and 82.5–87.5 cents per hour. Comparison of the distribution of all workers with the distributions for skilled and semi-skilled workers would lead to the conclusion that it is the difference in wage rates paid for these two grades of skill that has caused this bi-modal appearance in the total distribution. This interpretation, however, should not be made without having examined the original data in much greater detail. It is necessary to know the relative number of workers of each grade of skill, the number of men and the number of women with the different rates paid to each for various operations, variations in wages paid in different parts of the country, variations in union and non-union shops, any subdivisions within the industry due to production of different types of products, and other possible factors that could not be foreseen. All of these points are discussed in the article from which these tables were taken.²

The industry was a small one, with 51 plants located in 21 states, at the time of the investigation. None of the plants employed over 500 persons. Practically all of the employees were men, 53.9 per cent of them being classified as skilled, 28.1 per cent as semi-skilled, and 18.0 per cent as unskilled. Because of the high degree of skill required, and the dangerous character of the work, wage rates were unusually high, but considerable dispersion in the wage level was found, even for the same grades of skill.

These differences seemed to have very little relation to geographical location of the plants. The lowest rates were paid by a southern plant, but other plants in the south paid higher rates than some in the north, and there was wide variation even within single states. Union organization was also found to be a negligible factor—only 4 plants with 156 workers had collective agreements with trade unions.

Some relation appeared to exist between differences in wage scales and the two types of explosives, which were usually manufactured in separate plants. High explosives, dynamite, T.N.T., etc., were made by 32 of the plants, and black powder by 19. The latter product in-

² *Monthly Labor Review*, Vol. 47, No. 2 (August, 1938), pp. 378–94.

volves mechanical processes almost entirely, and the work is slightly less dangerous and less skilled than in the manufacture of high explosives. A distribution of average wages according to specific processes indicated that:

The averages of the two skilled occupations peculiar to black-powder manufacture were 76.4 cents for black-powder maker operators and 68.4 cents for black-powder line operators, both of the averages being lower than the lowest-paid skilled occupation peculiar to the high explosive branch (76.6 cents, for dope-house operators). The packers, found both in high explosive and in black-powder plants, averaged 74.5 cents.

As regards the semiskilled workers, however, it will be seen that the high explosive helpers averaged somewhat less than the helpers in black-powder plants.

Another factor in wage variation proved to be more significant than type of product. It was found that the 51 plants were controlled by 19 companies, the largest of which were known as the "Big Three."

The "Big Three" predominated more in the high-explosives branch of the industry; they controlled 18 plants with 2,394 wage earners, the "Other companies" had 14 establishments with 664 workers. In the black-powder branch, on the other hand, the "Big Three" had 10 establishments with 387 workers and the "Other companies" had 9 plants with 369 wage earners.

There was a difference in the average hourly earnings of all workers in black powder (72.1 cents) and high explosives (78.4 cents). However, the figures for the "Other companies" show higher earnings in black powder (67.0 cents) than in high explosives (64.3 cents). Among the "Big Three," this relationship is reversed, the respective averages being 77.4 and 82.7 cents. As indicated, the "Big Three" predominate more in the high explosives than in the black-powder branch, thus accounting for the higher average in the former as compared with the latter for all establishments. In other words, the industry-wide averages appear to reflect corporate wage policy rather than differences in the two branches.

Another table in the article separates the wage rates into two distributions, for the "Big Three" and for all the other smaller companies, each subdivided according to the three grades of skill. This table shows that the bi-modal flat-topped distribution of wage rates in the industry as a whole was due to the different wage scales paid by the "Big Three" and by the "Other companies." In the "Big Three" the modal rate was 82.5–87.5 cents per hour, while in "Other companies" it was 62.5–67.5, both skilled and semi-skilled centering at the mode in the

latter case. A much smaller differential between the wages paid to skilled labor and those paid to semi-skilled labor existed in the "Other companies" than in the "Big Three."

Less variation was apparent among the individual concerns in the "Other companies" group than among those constituting the "Big Three."

Among the small companies, each of the distributions [according to skill] had essentially a single mode. Each of the distributions of the "Big Three" had several modes, reflecting further marked differences in the wage policies of these companies, which were alike only in the fact that their wages were markedly higher than those of the small companies.

Differences among the several plants controlled by each company in the "Big Three" are also noted.

No data are given to show the proportions of skilled and of semi-skilled workers employed in manufacturing the two kinds of explosives, nor the difference between the rates paid to these workers respectively in the two types of companies. Thus the interpretation of the data remains incomplete. It can only be assumed that the higher average rates paid by "Other companies" to black-powder workers must be the result of special conditions in a few individual plants.

Arrangement

It might appear that any report would automatically follow the chronological procedure of the investigation. In practice, however, few reports follow the exact sequence of events as they occurred.

The writing must be organized so as to attract the attention of the reader. This can usually be accomplished by presenting a summary of the findings of the study immediately following the statement of the problem. The most important features of the work should follow the summary, leaving less important parts to a later point. Any outstanding feature such as an unusual problem of collection, a particularly striking summary tabulation, a penetrating technique of analysis or the discovery of an important new relationship, should be given prominence early in the report. The further organization might place the results of analysis next followed by a description of the collection process and a statement of the conclusions, with the details of analysis in an appendix. An inverted arrangement of this kind may be justified as a means of securing emphasis.

In stressing the particularly significant parts of the work, the general arrangement must not be disregarded. The report must possess continuity or the reader will lose interest. The achievement of an integrated whole may require an increased amount of connective writing when the parts are presented out of chronological sequence, but this is a desirable addition in producing an effective report.

Style

Style of writing is as much an individual matter as finger prints. For that reason each person must develop his own style, but there are certain fundamentals which form a necessary common basis for this development.

Clarity.—Clear statement proceeds from clarity of ideas. The author must be thoroughly familiar with his material and then must set for himself the task of presenting it lucidly. His writing must clarify and not mystify. Clarity is seldom achieved even by experienced writers in the first draft. Putting one's ideas on paper is only the first step in their logical arrangement, but a critical and objective rereading of this first draft will often indicate to the writer what he is really trying to say, and how it may be expressed more simply and clearly.

Directness.—There is no place in the preparation of reports for ornamental writing. The goal is effective presentation through direct statement.

Precision.—Many reports lose effectiveness through loose construction and lack of precision in the use of words. Such writing gives the same impression that one obtains by looking at a landscape through binoculars that are out of focus. Some examples culled from student reports illustrate lack of precision. In each example a corrected statement follows the quoted original.

Example 1: Discussion of a graph containing per capita consumption of lamb and the average price of lamb in the United States annually 1917–1936:

Student: The graph may serve to show the determination of prices as a supply and demand schedule. It is clear that prices were lowered with an effort to stimulate the per capita consumption. It would mean that the farmers would have to raise more to earn the same amount of money than before the lowering of prices but it would create a demand that would enable them to again raise the prices with a slight reduction of produce and make the same profits.

Corrected: The graph may be considered as a series of demand-price relations. The demand for lamb apparently is fairly elastic because proportionally smaller decreases in price accompany increases in per capita consumption and vice versa. Hence lower prices mean that farmers will have to market more lambs to maintain their total income from this source. It is possible, however, that some of the augmented demand stimulated by lower prices will be retained in a subsequent period of higher prices so that surpluses can be avoided and profits can be maintained.

Example 2: Discussion of a table containing sales of retail chain stores in the United States for the months of September, October, and November, 1936:

Student: The November figures give every indication of the December business reaching a very high level. This is due to two reasons namely the November figures show a very small decrease in comparison with the October figures which are the peak of the fall season. Therefore due to the Christmas season, and due to the small increase in November, everything seems to indicate that December's business will reach a new high in comparison with the December's of the depression years.

Corrected: Sales of chain stores are usually higher in October than in November but the very slight decline in November this year suggests that December sales may reach a fairly high level.

Example 3: Discussion of table showing changes in the ratings from 1927 to 1932 of railroad bonds listed in Moody's *Manual of Investments*:

Student: In group III, 27 out of 38 bonds decreased in rating by one point, the other 11 remaining the same. We thus conclude that over the five-year period there was a slight but decidedly large number of bonds that dropped in value.

Corrected: Of the 38 bonds in group III in 1927, 11 retained the same rating in 1932 while 27 were rated one grade lower. We conclude, therefore, that in the five-year interval during which business changed from prosperity to the depth of depression none of these bonds suffered a major loss of investment position, but that a large number of them were considered to be somewhat less desirable as investments.

Avoidance of Repetition.—A common fault of students is repetitious writing. This can be eliminated in the long run by correcting such basic faults as lack of concentration, incomplete or defective outlining, and failure to visualize the subject clearly. Much of the repetition could be detected and eliminated even by an inexperienced writer

through careful reading and correction of his own finished product. Many reports presented to teachers bear the unmistakable evidence of never having been read after they were written. A great gain has been made when a student finally abandons the illusion that one statement repeated in three different forms will be accepted by the instructor as three different ideas.

Appearance

Part of the impression made by a report depends upon its appearance. Consequently attention must be paid to form, neatness, penmanship or typing, margins, and similar mechanical features. The planning and execution of tables and charts are particularly important because the impression they will make on the reader depends entirely on their appearance. Pages should be numbered and only one side of the paper should be used.

THE FORM OF A REPORT

Each report is an individual piece of work having a definite purpose, and as such it should be prepared in the manner which will be most effective. There are, however, a few generally accepted rules of form which provide a stable basis on which to build individual differences. These general rules relate to the basic structure of presentation and have sometimes been referred to as the anatomy of the report.

Title Page

The title page should give a brief name of the report, its subtitle if any, the author, the date, and any further identification which is pertinent. The wording which appears on the title page is usually not a complete statement of the problem to be presented. For example the title may be "Foreign Trade of the United States" whereas the subject for investigation may be "The declining share of the United States in the total value of world trade studied separately for imports and exports annually since 1928."

The title page of an unpublished report is similar to that of a book except for the publishing agency. Students should consult the title pages of several books or pamphlets in order to familiarize themselves with the usual form.

Statement of Problem

If the subject of the problem has been assigned to the investigator in writing, the original or a copy of it should be included with his report. This serves two purposes: it shows the authorization for the work, and provides a complete statement of the problem for the reader. If the problem has been amended in any way during the investigation, such alterations should be stated at this point.

Sometimes there are special circumstances connected with the problem which have determined its limitations and which should be understood by the reader before he reads the report. All such explanations should be brought in along with the statement of the problem. They can be thought of as a foreword and in a report of some length grow into a preface.

Table of Contents

A book is usually organized by sections and chapters, and the list of these with the page reference for each appears as the table of contents. The same plan should ordinarily be followed in a report. This gives the reader an outline in sufficient detail to show the order of development and the importance of the several parts of the report.

List of Illustrations

If a report contains several tables, graphs, or other exhibits these may be combined in a single list of illustrations. At the discretion of the writer separate lists for each kind of illustration may be substituted for the combined list. A single table or graph appearing in a report might be referred to at the end of the table of contents, it might be given separate listing, or reference to it might be omitted.

The Body of the Report

The greater part of any report is devoted to the exposition of the problem itself and its solution. This consists of three parts: introduction, main text, and conclusions.

Introduction.—The purpose and plan of the research work should be explained in sufficient detail so that the reader can visualize the entire problem. The length of this statement will depend upon the size of the complete report and the amount of new material which it contains. In advanced research work twenty or more pages of intro-

ductory statement may be required, but in the usual student report a simple statement of scope and purpose is all that will be needed.

These first paragraphs contain the author's real introduction to his reader. The impression they make will determine to a large extent the attitude of the reader toward the entire report. Great care should be exercised, therefore, to make the introductory statement as effective as possible.

Main Text.—This section must contain an account of the entire procedure of the investigation including a description of the collection process, a statement of the method of analysis, an explanation of all tables and graphs, and any other parts of the development of the subject. In order to keep before the reader the plan of organization of the report it is desirable to use headings and subheadings throughout. These aid in giving proper emphasis to major divisions of the work and prevent overemphasis of subordinate parts.

The collection of data should be described. If the data have been obtained from direct sources, a complete presentation of the collection procedure will be required. If library sources have been used, it will be sufficient to give references along with the tables, explanations being confined to anything of an unusual nature encountered during the collection process.

The method of analysis should be set forth in advance of the presentation of the actual data employed. In many cases alternative methods of analysis are available. Under such circumstances the investigator should state why certain techniques were used and others rejected. This applies to graphs as well as to computed figures such as averages, index numbers, and coefficients.

In general, tables, graphs, or other exhibits relating to the collection or analysis of data should be inserted at the point at which they enter the development of the report. A combination of tables and graphs in an appendix and explanations of them in the main text is inconvenient because of the amount of leafing back and forth that is required. The maximum ease of reading results from placing tables and graphs on pages facing their descriptions. When a graph and its accompanying table appear in the text of the report, the two should be placed on facing pages. The discussion may then appear on the same page with the table or on an adjacent page.

Often something of particular significance will appear during the analysis. If this is not an integral part of the main subject matter but

some collateral information, it can be made a separate part of the report so as not to interrupt the principal development. In general anything which does not contribute directly to the major subject matter but which is considered sufficiently important to be included in the report should be given a separate place in the body of the report or should be placed in the appendix.

Conclusions.—As previously stated the conclusions may be placed at the beginning of a report for the sake of emphasis. They may appear in natural sequence at the end of the report. More commonly they are scattered through the report. That is, each part of the subject is developed completely in its place including the interpretation of the results of the analysis. When this is done it is desirable to bring the several partial conclusions together at the end in a final summary.

Appendix

Reference has been made at several points in the preceding discussion to material which may be put in an appendix. In each case the decision rests on whether the information is essential to the major development of the report or stands in collateral relation. The usual practice is to remove from the main text to an appendix long primary tables, detailed computations, collection forms, or similar materials which provide essential background for the report but are not integral parts of the major development.

Bibliography

A list of all published material consulted in the course of the investigation should be placed at the end of the report. This bibliography does not replace specific references made in the body of the report, but merely brings together in one place all reference material used, whether previously cited or not.

PROBLEMS

1. Why is the development of ability to write a well-organized report in good clear English an important part of the training of a statistician?
2. Should an outline for a report be drawn up before or after the report is written? Discuss your own procedure in writing a term report, and defend its merits, if any.

3. Discuss two or more possible types of readers to whom reports on each of the following subjects might be addressed. In each case suggest differences in content, style of writing, form, etc., that would make the reports suitable for the several kinds of readers.
 - (1) Causes of traffic accidents.
 - (2) The possibility of reducing costs of housing by introducing less durable plumbing fixtures.
 - (3) Need for aluminum conservation for defense purposes.
 - (4) Who are the buyers of used cars?
 - (5) Charge accounts vs. installment buying in department stores.
 - (6) Reading preferences of people who ride on the subway.
 - (7) Moving-picture attendance by children under 16.
 - (8) The effect on sales of changes in type of package and the use of color in packaging.
4. The instructor will assign one of the following subjects to each student for a written term report. Similar timely subjects should be added to the list.
 - (1) Study the relationship between flour milling in Buffalo, in Kansas City, and in Minneapolis, annually 1917-38. What percentage of the total flour that was milled annually in the United States was milled in each of these three cities during the same period? Interpret your answer.
 - (2) Construct an index of the amount of dividends paid by a fixed list of 10 light and power companies in the United States from 1929 to date. Compare this index with an index of payrolls for the light and power industry.
 - (3) Make a study of Woolworth, Murphy, and Kresge stores annually, 1929 to date, as to: total sales, sales per store, total profits, profits per store.
 - (4) Compare the consumption of cotton in manufacturing establishments in the New England States and in the Southern States, by decades 1840-1900, and for 1904, 1909, 1914, 1919, 1925, 1929, 1933, and 1937. Compare the number of establishments in the same sections for the same period.
 - (5) Select 5 large steel companies. Study the annual changes in sales, net earnings, and dividends paid from 1929 to date.
 - (6) Compare the sales by states for the most recent years available, of Ford, Chevrolet, and Plymouth cars, and interpret. What percentage were these three of total car sales in each state? (See statistical number of *Automobile Topics*, usually in February.)
 - (7) Is the price of steel scrap a forecaster of stock market prices in general? Study monthly figures for the past six years.

- (8) Multiply an index of industrial production by an index of commodity prices and divide the results by an index of department store sales. Compare this ratio with the Annalist Index of Business Activity (monthly, 1935, to date). If the ratio is assumed to measure maladjustment between production and distribution of goods, what is the relation of such maladjustment to the expansion and contraction of business activity?
- (9) Study the trend of the population from farm to city and from city to farm annually since 1920, and interpret.
- (10) Study the effect of the growth of trucking and the increased canal and lake traffic on railway freight over the past 15 years.
- (11) Study the changes in exports and imports of leading food products by the United States in 1929, 1934, and 1939.
- (12) Compare the number and per cent gainfully employed of the male and female population, by five-year age groups, for 1930 and 1940.
- (13) Discuss the relationship of changes since 1930 in Japan's total exports and her exports to the United States, and in her total imports and imports from the United States, naming the most important specific commodities involved.
- (14) As of 1929 and the most recent *Census of Manufactures*, compare the cost of raw materials, the value added by manufacture and the relation of wages to value added by manufacture, for the steel industry (not including blast furnaces) and for the cotton textile industry.
- (15) For the most recent date available, determine how many hours a road laborer in Boston, Massachusetts and in Savannah, Georgia, must work in order to buy the following supplies:

1 lb. bacon	7 qts. milk
3 lbs. beef	1 lb. coffee
12 lbs. bread	1 lb. butter
2 cans tomatoes	1 doz. eggs
5 lbs. sugar	
- (16) Make a study of significant shifts in the use of advertising media for last year as compared with the preceding year, including discussion of the apparent trend during the most recent months available.
- (17) Compare the number of man-days lost through strikes in 1941 with the highest and lowest years in the periods 1927-32 and 1933-40.
- (18) Prepare a graph showing production, shipments, and stocks on hand of Portland cement from 1920 to date (monthly data). What indication do you find of changes in production planning?
- (19) For the most recent year available and for a period about 10 or 15 years ago, compare the per capita average earnings of factory workers' families with the per capita cash income of farm families from farm marketings. Taking into account the probable additional income in

each type of family, the different standards of living and the expenses of each, do you conclude that the farmers are in need of continued government aid?

- (20) The 1940 annual statistical report of a wholesale grocery association with a membership of 24 wholesale grocery companies gives the information in column 1.

	(1) AVERAGE FOR 24 COMPANIES	(2) COMPANY X
Gross sales	\$640,000	\$200,000
Cost of goods sold.....	400,000	110,000
Gross margin	240,000	90,000
Distribution of gross margin in percentages		
Selling cost	40	37
Delivery, truck, and freight.....	12	18
Warehouse costs	8	5
Overhead	31	36
Profit	9	4
	100	100

You are employed as research clerk in the treasurer's office of Company X (a member of the association). In that capacity you have compiled the figures given in column 2. Write a report interpreting this statement for the president of your company.

REFERENCES

EIGELBERNER, J., *The Investigation of Business Problems*. Chicago and New York: A. W. Shaw Co., 1926.

Chapters XV, XVI, and XVII contain suggestions for the development of hypotheses in analytical investigation and the interpretation of the results of analysis. Chapter XVIII deals with report writing.

Manual on Research and Reports, issued by the Committee on Research of the Amos Tuck School of Administration and Finance, Dartmouth College. Baltimore: The Williams and Wilkins Co., 1933.

An outline of research procedure and a complete statement of the fundamentals of report writing.

SAUNDERS, ALTA G., and ANDERSON, CHESTER R., *Business Reports*. New York: McGraw-Hill Book Co., Inc., 1929.

Chapters XII to XVII should be required reading for every potential report writer.

SCHLUTER, WILLIAM C., *How To Do Research Work*. New York: Prentice-Hall, Inc., 1929.

The author's point of view concerning report writing makes chapters XIV-XVI well worth reading.

SPAHR, WALTER E., and SWENSON, RINEHART J., *Methods and Status of Scientific Research*. New York: Harper and Bros., 1930.

Chapter XII deals directly with the preparation of a report, but chapter XI on analysis and interpretation and chapter XIII on grammar, composition, and form will repay careful reading.

Standards of Research. Des Moines: Meredith Publishing Co., 1929.

The essentials of report writing are outlined in pages 49 to 54.

APPENDIX C

LOGARITHMS OF NUMBERS

STATISTICIANS employ many devices to lighten the inevitable burden of computation that accompanies analysis of numerical data. One such device is the use of logarithms. Their use simplifies the operations of multiplication, division, and raising to powers and is indispensable in extracting roots.

The theory of logarithms involves more advanced mathematical knowledge than the user of this book is assumed to possess. Fortunately the use of logarithms can be explained without extensive reference to the background of theory. The present exposition therefore is confined to the minimum essentials for effective use of logarithms as a tool of calculation.

DEFINITION

A definition can be developed by proceeding from known relations to new ones. The equalities on the left of the following list are well known.

$10,000 = 10^4$	$\log 10,000 = 4$
$1,000 = 10^3$	$\log 1,000 = 3$
$100 = 10^2$	$\log 100 = 2$
$10 = 10^1$	$\log 10 = 1$
$1 = 10^0$	$\log 1 = 0$
$.1 = 10^{-1}$	$\log .1 = -1$
$.01 = 10^{-2}$	$\log .01 = -2$
$.001 = 10^{-3}$	$\log .001 = -3$
$.0001 = 10^{-4}$	$\log .0001 = -4$

This is simply a list of the values of the successive powers of ten running from +4 to -4. The list might be extended indefinitely in each direction, but those given will be sufficient for our purposes. It will be noted for subsequent use that the powers of ten read from the top of the list down are in arithmetic progression, while the numbers are in geometric progression with a common ratio of .1, i.e. $(10,000 \times .1) = 1,000$, $(1,000 \times .1) = 100$, etc.

The numbers, 10,000, 1,000, etc., are all expressed as powers of ten and this idea is all that is involved in a system of logarithms, namely, to express numbers as powers of ten instead of in the usual way. Hence, *the logarithm of a number is the power to which ten must be raised in order to produce the number.* The logarithmic statements at the right of the preceding list are merely another way of expressing the relations at the left. That is, the logarithm of 10,000 is 4 because ten¹ raised to the fourth power equals 10,000, and the logarithm of .01 is -2 because ten raised to the minus second power equals .01.

TWO PARTS OF A LOGARITHM

The list on page 845 contains numbers which are integral powers of 10; therefore their logarithms are all whole numbers. Suppose that the logarithm of 5,000 were wanted. That is, to what power must the base, 10, be raised in order to be equal to 5,000? Or stated in symbols, if $5,000 = 10^x$, what is the value of x ? Since $1,000 = 10^3$ and $10,000 = 10^4$, 10 raised to some power between the third and fourth will equal 5,000. Consequently x , the logarithm of 5,000, equals $3 + a$ fraction.

The value of x will contain a fraction for all numbers except integral powers of 10. Also the value of x will contain a whole number part for all numbers except those between 1 and 10. From the list on page 845 it can be seen that the logarithms of numbers

between 1,000 and 10,000	equal	$3 + a$ fraction
between 100 and 1,000	equal	$2 + a$ fraction
between 10 and 100	equal	$1 + a$ fraction
between 1 and 10	equal	$0 + a$ fraction
between .1 and 1	equal	$-1 + a$ fraction
between .01 and .1	equal	$-2 + a$ fraction
between .001 and .01	equal	$-3 + a$ fraction
between .0001 and .001	equal	$-4 + a$ fraction

The whole number part of a logarithm is called the *characteristic* and the fractional part is called the *mantissa*. Rules for determining the characteristic of the logarithm of a number are relatively simple and will be stated in the next section. The mantissa or fractional power of 10 cannot be computed by any simple process; hence logarithms would

¹ Any other number except 1 could be made the base of a system of logarithms, but only the Common or Briggsian System using 10 as the base is employed in applied statistical work.

be of little practical value except for the fact that tables of mantissas have been prepared. The values appearing in a table are obtained by solution of an infinite series known as the logarithmic expansion. It is unnecessary to reproduce the expansion; since this theoretical background is not essential to ability to use the tables.

Rules for Characteristics

The tables do not contain characteristics; therefore rules must be developed for prefixing the whole number part of a logarithm to the fractional part found in the table. These rules must be so stated that both the sign and the numerical value of the characteristic can be determined. For this purpose all positive numbers are divided into two groups, (1) numbers greater than 1, (2) numbers between 0 and 1. The reason for this division will be clear from the list on page 845. The logarithm of 1 is 0 and the logarithms of all the numbers in the list greater than 1 are positive while the logarithms of all the numbers in the list less than 1 are negative. But these numbers are all integral powers of 10 and the mantissas of their logarithms are 0. Hence the observed relation becomes the first part of the rules for determining characteristics.

The second part of the rules is deduced from study of the list on the preceding page. The logarithms of numbers between 1,000 and 10,000 are $3 + \text{a fraction}$ and all such numbers have *four digits to the left of the decimal point*. The logarithms of numbers between 100 and 1,000 are $2 + \text{a fraction}$ and all such numbers have *three digits to the left of the decimal point*. The logarithms of numbers between 10 and 100 are $1 + \text{a fraction}$ and all such numbers have *two digits to the left of the decimal point*. The logarithms of numbers between 1 and 10 are $0 + \text{a fraction}$ and all such numbers have *one digit to the left of the decimal point*. The characteristics of the logarithms of all numbers between 1 and 10,000 are numerically one less than the number of digits to the left of the decimal point in the numbers themselves and the same can be inferred for all numbers greater than 10,000. Therefore we are ready to state:

Rule 1: The characteristics of the logarithms of all numbers greater than one are positive and their numerical values are one unit less than the number of digits to the left of the decimal point in the numbers themselves.

Examples using Rule I:

NUMBER	CHARACTERISTIC OF LOGARITHM
286.	2
12,769.	4
1,008.73	3
1.827	0
4,729,246.	6
387.094	2
42,700.892	4
10,000,000.	7

The same type of deduction leads to the rule for characteristics of the logarithms of numbers between zero and one. The logarithms of numbers between 1 and .1 are $-1 + \text{a fraction}$ and all such numbers have no 0's between the decimal point and the first significant digit. The logarithms of numbers between .1 and .01 are $-2 + \text{a fraction}$ and all such numbers have one 0 between the decimal point and the first significant digit and so on. Hence we can infer:

Rule II: The characteristics of the logarithms of all numbers between zero and one are negative and their numerical values are one unit greater than the number of zeros between the decimal point and the first significant digit of the numbers themselves.

Examples using Rule II:

NUMBER	CHARACTERISTIC OF LOGARITHM
.764	-1 or 9. -10
.031	-2 8. -10
.02793	-2 8. -10
.00004	-5 5. -10
.01004	-2 8. -10
.80086	-1 9. -10
.8	-1 9. -10
.0000001	-7 3. -10

Negative characteristics are bothersome in computations; therefore the usual procedure is to write them in the form indicated at the right of the preceding examples.

The two rules for characteristics cover all values from $-\infty$ to $+\infty$. That is, the logarithms of numbers greater than 1 may have characteristics ranging from 0 to $+\infty$ and those of numbers between 0 and 1 may have characteristics ranging from -1 to $-\infty$. Therefore all possible characteristics apply to numbers greater than 0, and there is no such thing as the 'logarithm of a negative number or of 0. This is also evident from the fact that any positive or negative power of a positive number (the base 10 in this case) is a positive number.

Use of Table of Mantissas

A table of mantissas, more commonly referred to as a logarithm table, is merely a listing of numbers and the fractional powers of 10 which will produce the numbers. As previously stated it is this table of fractional powers which gives value to the logarithmic device as a practical tool. Hence familiarity with the use of the table is a primary requirement for the user of logarithms.

The table on pages 857-74 is a five-place table, i.e., the powers of 10 are carried to five decimal places. The table contains no characteristics. They are to be supplied by the use of the two rules already quoted. The five places in the table are to be read in all cases with a decimal point preceding.

These decimal fractions are all positive; therefore the user must be cautious on two points (1) that the prefixing of a negative characteristic does not lead to the error of treating the mantissa as negative and (2) that in subtracting a larger logarithm from a smaller one the actual process is not reversed so as to give a negative mantissa which cannot be found in the table. The first of these difficulties is overcome in practice by writing negative characteristics in the form indicated in the list on page 848. The method of avoiding the second is illustrated at the top of page 853, and also in several steps of the examples on page 854.

Reading Logarithms of Numbers.—The mantissa of any number of not more than four digits can be read directly from the table. The first three digits are found in the column labeled "N" at the left of the page and the fourth digit is found in the row comprising the column headings at the top of the page. The mantissa of any number is the number at the intersection of the row containing the first three digits of the number at the left and the column containing the fourth digit at the top. Thus to find $\log 4273$ turn to page 863 and find 427 at the left, then find 3 at the top of the page and at the intersection of this row and column find the mantissa .63073. The number has four digits to the left of the decimal point; therefore the characteristic is three and $\log 4273 = 3.63073$, or $4273 = 10^3 \cdot 63073$.

It will be noted that the use of the table would have been identical if the number had been 42.73 or any other number containing the consecutive digits 4, 2, 7, and 3. The logarithms of such numbers differ only in their characteristics. Thus,

$\log 427300 = 5.63073$	$\log 42.73 = 1.63073$
$\log 42730 = 4.63073$	$\log 4.273 = .63073$
$\log 4273 = 3.63073$	$\log .4273 = 9.63073 - 10$
$\log 427.3 = 2.63073$	$\log .04273 = 8.63073 - 10$

The logarithms in the following list should be verified for practice.

$\log 1809 = 3.25744$	$\log .05188 = 8.71500 - 10$
$\log 4.555 = 0.65849$	$\log .98 = 9.99123 - 10$

The table is constructed so that the last three digits of the mantissa are found at the intersection of the proper row and column and the first two digits are found by glancing up the same column until figures in the first two positions are printed. A table prepared in this way is more readable than one in which the five digits of each mantissa are printed. When the second digit increases by 1, all five digits of ten succeeding mantissas are printed and the five digits are printed in the first row of each page. With this plan the first figures that meet the eye as it moves up the column will always be the correct ones for the first two digits unless, of course, all five digits happen to be printed at the intersection of the row and column in which the mantissa of the given number is found.

The mantissas of numbers containing more than four digits are found with the aid of the proportional parts section at the right of each page of the table. The mantissa corresponding to the first four digits of a number is read in the usual way. Then the eye should carry over to the proportional parts section in the same row in which the mantissa of the first four digits has been found. The proportional part in this row at the intersection of the column containing the fifth digit of the number at the top of the page is the amount which must be added to the mantissa of the four-digit number to give the mantissa of the five-digit number.

To find $\log 12386$

$\log 12380$	$= 4.09272$
proportional part for 6 in the fifth place	$= 21$
$\log 12386$	$= 4.09293$

The logarithm of a six-digit number can also be found by using the proportional parts section successively and moving the proportional part one place to the right for the sixth digit of the number.

To find $\log 22.0937$

$$\begin{array}{rcl}
 \log 22.09 & & = 1.34420 \\
 \text{proportional part for 3 in fifth place} & = & 6 \\
 \text{proportional part for 7 in sixth place} & = & 14 \\
 \hline
 \log 22.0937 & & = 1.34427
 \end{array}$$

Reading Antilogarithms.—The use of logarithms for calculation requires that the number corresponding to a given logarithm be read from the table. This is the reverse operation of that explained in the preceding section; hence the process involves finding the given mantissa in the table and reading the corresponding number at the left and top.

$$\begin{array}{ll}
 \text{If } \log x = 2.24080, & x = 174.1 \\
 \log x = 0.89900, & x = 7.925 \\
 \log x = 7.74663 - 10, & x = .00558
 \end{array}$$

In determining the position of the decimal point in the antilogarithm the rules for characteristics are used in reverse form. Thus in the first example we say "the characteristic is positive; therefore the number is greater than one, and since the characteristic is 2, the number has one more than 2, or 3 digits to the left of the decimal point." In the third example the characteristic is negative; hence the number is less than one and since the characteristic is -3 , the number has one less than 3, or 2 ciphers between the decimal point and the first significant figure.

The proportional parts section is used to find the antilogarithm when the mantissa of the given logarithm is not found exactly in the table. The first four digits of the antilogarithm are taken corresponding to the mantissa in the table nearest to but less than the given mantissa. The difference between the mantissa from the table and the given mantissa is found in the same row of the proportional parts section and the figure at the top of this column is the fifth digit of the antilogarithm. To find the antilogarithm when $\log x = 3.67297$,

$$\log 4709 = 3.67293$$

hence the first four digits of the antilogarithm are 4709. Then $.67297 - .67293 = 4$, and the figure at the top of the column containing the proportional part 4 is 4. Therefore the antilogarithm of 3.67297 is 4709.4.

Antilogarithms to six digits can be found on the first few pages of the table where the proportional parts are large but not for the remainder of the table. Find to six digits the antilogarithm, if

$$\begin{array}{rcl} \log x & = & .03291 \\ \log 1.078 & = & .03262 \\ .03291 - .03262 & = & 29 \end{array}$$

The proportional part nearest to but less than 29 is 28; hence the fifth digit is 7. The difference ($29 - 28 = 1$) is considered as 10 in entering the proportional parts for the sixth digit. The proportional parts 8 and 12 are equidistant from 10 and the sixth digit is either 2 or 3. It could be taken as 25 in the sixth and seventh places. Therefore $\log 1.078725 = .03291$.

RULES FOR USING LOGARITHMS

Rule A: Multiplication

To obtain the product of two or more numbers add their logarithms and find the antilogarithm of the sum. In symbols the process is,

$$\log (a \times b \times c \cdots) = \log a + \log b + \log c + \cdots$$

and the antilogarithm of the sum at the right is the product of the numbers. The work is usually arranged as follows.

$$\begin{array}{l} \text{Example: } x = 369.8 \times 8007 \times .12366 \\ \log x = \log 369.8 + \log 8007 + \log .12366 \\ \log 369.8 = 2.56797 \\ \log 8007 = 3.90347 \\ \log .12366 = 9.09223 - 10 \\ \hline \log x = 15.56367 - 10 \\ \log x = 5.56367 \\ x = 366,160 \end{array}$$

Rule B: Division

To obtain the quotient of two numbers subtract their logarithms and find the antilogarithm of the difference. In symbols,

$$\log \left(\frac{a}{b} \right) = \log a - \log b$$

and the antilogarithm of the difference at the right is the quotient of the numbers.

$$\begin{aligned}
 \text{Example: } x &= \frac{7.4362}{10.82} \\
 \log x &= \log 7.4362 - \log 10.82 \\
 \log 7.4362 &= 10.87135 - 10 \quad (\text{add } 10. - 10 \text{ but never subtract} \\
 \log 10.82 &= 1.03423 \quad \quad \quad \text{the minuend from the subtrahend}) \\
 \log x &= 9.83712 - 10 \\
 x &= .68726
 \end{aligned}$$

Rule C: Raising to Powers

To obtain the value of a number raised to a power, multiply the logarithm of the number by the power and find the antilogarithm of the product. In symbols,

$$\log (a)^x = x \log a$$

and the antilogarithm of the product at the right is the value of the number raised to the power.

$$\begin{aligned}
 \text{Example:} \quad x &= (5)^6 \\
 \log x &= 6 \log 5 \\
 \log 5 &= .69897 \\
 &\quad \times 6 \\
 \hline
 \log x &= 4.19382 \\
 x &= 15625
 \end{aligned}$$

Rule D: Extracting Roots

To obtain the value of the root of a number divide the logarithm of the number by the root and find the antilogarithm of the quotient. In symbols,

$$\log \sqrt[n]{a} = \frac{\log a}{n}$$

and the antilogarithm of the quotient at the right is the value of the root of the number.

$$\begin{aligned}
 \text{Example:} \quad x &= \sqrt[3]{.027416} \\
 \log x &= \frac{\log .027416}{3}
 \end{aligned}$$

$$\begin{array}{rcl}
 \log .027416 & = & 8.43801 - 10 \\
 & \begin{array}{r} 20 \qquad - 20 \\ 3 \overline{) 28.43801 - 30} \end{array} & \text{(add and subtract enough} \\
 & & \text{to give } -10 \text{ at the right} \\
 \log x & = & 9.47934 - 10 \text{ after dividing by 3)} \\
 x & = & .30154
 \end{array}$$

Combinations

The usual problem employing logarithms makes use of two or more of the foregoing rules. An example with symbols will show how the several rules are applied, taking one step at a time.

$$x = \sqrt[n]{a^k \times b \over \sqrt[p]{d}}$$

$$\log x = \frac{1}{n} \log \left[\frac{a^k \times b}{\sqrt[p]{d}} \right] \quad \text{Rule D}$$

$$\log x = \frac{1}{n} [\log(a^k \times b) - \log \sqrt[p]{d}] \quad \text{Rule B}$$

$$\log x = \frac{1}{n} [\log a^k + \log b - \log \sqrt[p]{d}] \quad \text{Rule A}$$

$$\log x = \frac{1}{n} \left[k \log a + \log b - \frac{1}{p} \log d \right] \quad \text{Rules C and D}$$

In actual practice, all of the conversion to the form for logarithmic computation should be carried out in a single step, as follows.

Example:

$$x = \sqrt[4]{(386)^2 \times 47.21 \times \sqrt{.086} \over \sqrt[3]{2.1744 \times (4)^4}}$$

$$\begin{array}{rcl}
 2 \log .386 & = & 9.17318 - 10 \\
 \log 47.21 & = & 1.67403 \\
 \frac{1}{2} \log .086 & = & 9.46725 - 10 \\
 \hline
 & & 10.31446 - 10
 \end{array} \quad (1)$$

$$\begin{array}{rcl}
 \frac{1}{8} \log 2.1744 & = & .04217 \\
 3 \log 4 & = & 1.80618 \\
 \hline
 & & 1.84835
 \end{array} \quad (2)$$

$$(1) - (2)$$

$$\begin{array}{rcl}
 & & 8.46611 - 10 \\
 & & 30. \qquad - 30 \\
 & & \hline
 & & 4) 38.46611 - 40 \\
 \log x & = & 9.61653 - 10 \\
 x & = & .41355
 \end{array}$$

Example: $x = \sqrt[10]{365 \times (400)^5 \times (425)^2 \times 512 \times 603}$

$$\begin{array}{rcl} \log 365 & = & 2.56229 \\ 5 \log 400 & = & 13.01030 \\ 2 \log 425 & = & 5.25678 \\ \log 512 & = & 2.70927 \\ \log 603 & = & 2.78032 \\ 10) \overline{26.31896} & & \\ \log x & = & 2.63190 \\ x & = & 428.45 \end{array}$$

This example shows the method of finding the geometric average of a set of numbers.

PROBLEMS

1. Find the logarithms,

a) 8654	e) .0006
b) 37.29	f) 1332
c) 6060	g) .998
d) .0731	h) 1,394,000

2. Find the logarithms,

a) 12.607	d) 100625
b) .0072162	e) 865094
c) 586420	f) .931442

3. Find x in each of the following,

a) $\log x = 3.57229$	d) $\log x = 2.14270 - 10$
b) $\log x = 9.83283 - 10$	e) $\log x = 6.00130$
c) $\log x = .20249$	f) $\log x = 7.99003 - 10$

4. Find x in each of the following,

a) $\log x = 3.99006$	c) $\log x = .17700$
b) $\log x = 7.00009 - 10$	d) $\log x = 4.13713$

5. Find x by logarithms,

a) $(10)^3 \times .0001 \times \sqrt[4]{10000}$

b) $\frac{(1)^5 \times \sqrt{10000}}{1000000}$

c) $\sqrt[4]{\frac{10000000 \times (.1)^4}{\sqrt[5]{.00001}}}$

d) $\sqrt[6]{\frac{(.01)^5 \times \sqrt[3]{1000000}}{(10)^4}}$

6. Find the square root of each of the following numbers,
 - a) 267.09
 - b) 36412.4
 - c) 4096
 - d) 728.97
 - e) 9801.01
 - f) 9.80101
 - g) .010406
 - h) .10406
7. Find the cube root of each of the following numbers,
 - a) 4000000
 - b) .000949
 - c) .00949
 - d) .0949
8. In 1937, 873,993,000 bushels of wheat were grown on 64,460,000 acres of land in the United States. What was the average yield per acre?
9. The farm value of the 1937 wheat crop of Problem 8 was \$869,140,000. Find the average farm price per bushel and the average return to the farmer per acre of wheat harvested.
10. The number of passenger cars registered in the United States in 1938 was 25,261,649. Assuming an average value of \$435.25 per car, find the total value of the passenger cars in use.
11. 369,475 pounds of coal were sold for \$4.75 per long ton. Find the income from the sale.
12. The trend of a series is a compound interest curve increasing at the rate of 1.375 per cent per year. If the value of the trend is 16427 in 1920, what will its value be in 1942? That is, find A in the equation: $A = 16427 (1.01375)^{22}$.
13. Find the value of the 20th term in the geometric series 2, 4, 8, 16, 32 (In a geometric progression $t_n = t_1 r^{n-1}$. See footnote 12, page 337.)
14. The sales of a department store increased from \$780,000 in 1930 to \$3,725,000 in 1940. What was the average annual rate of growth?

Five-Place Logarithms of Numbers

100 - 149

N	0	1	2	3	4	5	6	7	8	9	Proportional Parts								
											1	2	3	4	5	6	7	8	9
100	00000	00043	00087	00130	00173	00217	00260	00303	00346	00389	4	9	13	17	22	26	30	34	39
101	432	475	518	561	604	647	689	732	775	817	4	9	13	17	22	26	30	34	39
102	860	903	945	988	01030	01072	01115	01157	01199	01242	4	8	13	17	21	25	29	34	38
103	01284	01326	01368	01410	452	494	536	578	620	662	4	8	13	17	21	25	29	34	38
104	703	745	787	828	870	912	953	995	02036	02078	4	8	13	17	21	25	29	34	38
105	02119	02160	02202	02243	02284	02325	02366	02407	449	490	4	8	12	16	20	25	29	33	37
106	531	572	612	653	694	735	776	818	859	898	4	8	12	16	20	25	29	33	37
107	938	979	03019	03060	03100	03141	03181	03222	03262	03302	4	8	12	16	20	24	28	32	36
108	03342	03383	423	463	503	543	583	623	663	703	4	8	12	16	20	24	28	32	36
109	743	782	822	862	902	941	981	04021	04060	04100	4	8	12	16	20	24	28	32	36
110	04139	04179	04218	04258	04297	04336	04376	415	454	493	4	8	12	16	20	23	27	31	35
111	523	571	610	650	689	727	766	805	844	883	4	8	12	16	20	23	27	31	35
112	922	961	999	05038	05077	05115	05154	05192	05231	05269	4	8	12	16	20	23	27	31	35
113	05308	05346	05385	423	461	500	538	576	614	652	4	8	11	15	19	23	27	30	34
114	690	729	767	805	843	881	918	956	994	06032	4	8	11	15	19	23	27	30	34
115	06070	06108	06145	06183	06221	06258	06296	06333	06371	408	4	8	11	15	19	23	27	30	34
116	446	483	521	558	595	633	670	707	744	781	4	7	11	15	18	22	26	30	33
117	819	856	893	930	967	07004	07041	07078	07115	07151	4	7	11	15	18	22	26	30	33
118	07188	07225	07262	07298	07335	372	408	445	482	518	4	7	11	15	18	22	26	30	33
119	555	591	628	664	700	737	773	809	846	882	4	7	11	14	18	22	25	29	32
120	918	954	990	08027	08063	08099	08135	08171	08207	08243	4	7	11	14	18	22	25	29	32
121	08279	08314	08350	386	422	458	493	529	565	600	4	7	11	14	18	22	25	29	32
122	636	672	707	743	778	814	849	884	920	955	4	7	10	14	18	21	24	28	32
123	991	09026	09061	09096	09132	09167	09202	09237	09272	09307	4	7	10	14	18	21	24	28	32
124	09342	377	412	447	482	517	552	587	621	656	4	7	10	14	18	21	24	28	32
125	691	726	760	795	830	864	899	934	968	10003	4	7	10	14	18	21	24	28	32
126	10037	10072	10106	10140	10175	10209	10243	10278	10312	346	3	7	10	14	17	20	24	27	31
127	380	415	449	483	517	551	585	619	653	687	3	7	10	14	17	20	24	27	31
128	721	755	789	823	857	890	924	958	992	11025	3	7	10	14	17	20	24	27	31
129	11059	11093	11126	11160	11193	11227	11260	11294	11327	361	3	7	10	14	17	20	24	27	31
130	394	428	461	494	528	561	594	628	661	694	3	7	10	13	16	20	23	26	30
131	727	760	793	826	860	893	926	959	992	12024	3	7	10	13	16	20	23	26	30
132	12057	12090	12123	12156	12189	12222	12254	12287	12320	353	3	7	10	13	16	20	23	26	30
133	385	418	450	483	516	548	581	613	646	678	3	7	10	13	16	20	23	26	30
134	710	743	775	808	840	872	905	937	969	13001	3	6	10	13	16	19	22	26	29
135	13033	13066	13098	13130	13162	13194	13226	13258	13290	322	3	6	10	13	16	19	22	26	29
136	354	386	418	450	481	513	545	577	609	640	3	6	10	13	16	19	22	26	29
137	672	704	735	767	799	830	862	893	925	956	3	6	10	13	16	19	22	26	29
138	988	14019	14051	14082	14114	14145	14176	14208	14239	14270	3	6	9	12	16	19	22	25	28
139	14301	333	364	395	426	457	489	520	551	582	3	6	9	12	16	19	22	25	28
140	613	644	675	706	737	768	799	829	860	891	3	6	9	12	16	19	22	25	28
141	922	953	983	15014	15045	15076	15106	15137	15168	15198	3	6	9	12	16	19	22	25	28
142	15229	15259	15290	320	351	381	412	442	473	503	3	6	9	12	15	18	21	24	27
143	534	564	594	625	655	685	715	746	776	806	3	6	9	12	15	18	21	24	27
144	836	866	897	927	957	987	16017	16047	16077	16107	3	6	9	12	15	18	21	24	27
145	16137	16167	16197	16227	16256	16286	316	346	376	406	3	6	9	12	15	18	21	24	27
146	435	465	495	524	554	584	613	643	673	702	3	6	9	12	15	18	21	24	27
147	732	761	791	820	850	879	909	938	967	997	3	6	9	12	14	17	20	23	26
148	17026	17056	17085	17114	17143	17173	17202	17231	17260	17289	3	6	9	12	14	17	20	23	26
149	319	348	377	406	435	464	493	522	551	580	3	6	9	12	14	17	20	23	26

Five - Place Logarithms of Numbers

150 - 199

N	0	1	2	3	4	5	6	7	8	9	Proportional Parts								
											1	2	3	4	5	6	7	8	9
150	17609	17638	17667	17696	17725	17754	17782	17811	17840	17869	3	6	9	12	14	17	20	23	26
151	898	926	955	984	18013	18041	18070	18099	18127	18156	3	6	9	12	14	17	20	23	26
152	18184	18213	18241	18270	298	327	355	384	412	441	3	6	9	12	14	17	20	23	26
153	469	498	526	554	583	611	639	667	696	724	3	6	8	11	14	17	20	22	28
154	752	780	808	837	865	893	921	949	977	19005	3	6	8	11	14	17	20	22	28
155	19033	19061	19089	19117	19145	19173	19201	19229	19257	285	3	6	8	11	14	17	20	22	28
156	312	340	368	396	424	451	479	507	535	562	3	6	8	11	14	17	20	22	28
157	590	618	645	673	700	728	756	783	811	838	3	6	8	11	14	17	20	22	28
158	866	893	921	948	976	20003	20030	20058	20085	20112	3	5	8	11	14	16	19	22	24
159	20140	20167	20194	20222	20249	276	303	330	358	385	3	5	8	11	14	16	19	22	24
160	412	439	466	493	520	548	575	602	629	656	3	5	8	11	14	16	19	22	24
161	683	710	737	763	790	817	844	871	898	925	3	5	8	11	14	16	19	22	24
162	952	978	21005	21032	21059	21085	21112	21139	21165	21192	3	5	8	11	14	16	19	22	24
163	21219	21245	272	299	325	352	378	405	431	458	3	5	8	11	14	16	19	22	24
164	484	511	537	564	590	617	643	669	696	722	3	5	8	10	13	16	18	21	23
165	748	775	801	827	854	880	906	932	958	985	3	5	8	10	13	16	18	21	23
166	22011	22037	22063	22089	22115	22141	22168	22194	22220	22246	3	5	8	10	13	16	18	21	23
167	272	298	324	350	376	401	427	453	479	505	3	5	8	10	13	16	18	21	23
168	531	557	583	608	634	660	686	712	737	763	3	5	8	10	13	16	18	21	23
169	789	814	840	866	891	917	943	968	994	23019	3	5	8	10	13	16	18	21	23
170	23045	23070	23096	23121	23147	23172	23198	23223	23249	274	2	5	8	10	12	15	18	20	22
171	300	325	350	376	401	426	452	477	502	528	2	5	8	10	12	15	18	20	22
172	553	578	603	629	654	679	704	729	754	779	2	5	8	10	12	15	18	20	22
173	805	830	855	880	905	930	955	980	24005	24030	2	5	8	10	12	15	18	20	22
174	24055	24080	24105	24130	24155	24180	24204	24229	254	279	2	5	8	10	12	15	18	20	22
175	304	329	353	378	403	428	452	477	502	527	2	5	8	10	12	15	18	20	22
176	551	576	601	625	650	674	699	724	748	773	2	5	8	10	12	15	18	20	22
177	797	822	846	871	895	920	944	969	993	25018	2	5	8	10	12	15	18	20	22
178	25042	25066	25091	25115	25139	25164	25188	25212	25237	261	2	5	7	10	12	14	17	19	22
179	285	310	334	358	382	406	431	455	479	503	2	5	7	10	12	14	17	19	22
180	527	551	575	600	624	648	672	696	720	744	2	5	7	10	12	14	17	19	22
181	768	792	816	840	864	888	912	935	959	983	2	5	7	10	12	14	17	19	22
182	26007	26031	26055	26079	26102	26126	26150	26174	26198	26221	2	5	7	10	12	14	17	19	22
183	245	269	293	316	340	364	387	411	435	458	2	5	7	10	12	14	17	19	22
184	482	505	529	553	576	600	623	647	670	694	2	5	7	10	12	14	17	19	22
185	717	741	764	788	811	834	858	881	905	928	2	5	7	9	12	14	16	18	21
186	951	975	998	27021	27045	27068	27091	27114	27138	27161	2	5	7	9	12	14	16	18	21
187	27184	27207	27231	254	277	300	323	346	370	393	2	5	7	9	12	14	16	18	21
188	416	439	462	485	508	531	554	577	600	623	2	5	7	9	12	14	16	18	21
189	646	669	692	715	738	761	784	807	830	852	2	5	7	9	12	14	16	18	21
190	875	898	921	944	967	989	28012	28035	28058	28081	2	5	7	9	12	14	16	18	21
191	28103	28126	28149	28171	28194	28217	240	262	285	307	2	5	7	9	12	14	16	18	21
192	330	353	375	398	421	443	466	488	511	533	2	5	7	9	12	14	16	18	21
193	556	578	601	623	646	668	691	713	735	758	2	4	7	9	11	13	15	18	20
194	780	803	825	847	870	892	914	937	959	981	2	4	7	9	11	13	15	18	20
195	29003	29026	29048	29070	29092	29115	29137	29159	29181	29203	2	4	7	9	11	13	15	18	20
196	226	248	270	292	314	336	358	380	403	425	2	4	7	9	11	13	15	18	20
197	447	469	491	513	535	557	579	601	623	645	2	4	7	9	11	13	15	18	20
198	667	688	710	732	754	776	798	820	842	863	2	4	7	9	11	13	15	18	20
199	885	907	929	951	973	994	30016	30038	30060	30081	2	4	7	9	11	13	15	18	20

Five-Place Logarithms of Numbers

200 - 249

N	0	1	2	3	4	5	6	7	8	9	Proportional Parts								
											1	2	3	4	5	6	7	8	9
200	30103	30125	30146	30168	30190	30211	30233	30255	30276	30298	2	4	7	9	11	13	15	18	20
201	320	341	363	384	406	428	449	471	492	514	2	4	7	9	11	13	15	18	20
202	535	557	578	600	621	643	664	685	707	728	2	4	6	8	10	13	15	17	19
203	750	771	792	814	835	856	878	899	920	942	2	4	6	8	10	13	15	17	19
204	963	984	31006	31027	31048	31069	31091	31112	31133	31154	2	4	6	8	10	13	15	17	19
205	31175	31197	218	239	260	281	302	323	345	366	2	4	6	8	10	13	15	17	19
206	387	408	429	450	471	492	513	534	555	576	2	4	6	8	10	13	15	17	19
207	597	618	639	660	681	702	723	744	765	785	2	4	6	8	10	13	15	17	19
208	806	827	848	869	890	911	931	952	973	994	2	4	6	8	10	13	15	17	19
209	32015	32035	32056	32077	32098	32118	32139	32160	32181	32201	2	4	6	8	10	13	15	17	19
210	222	243	263	284	305	325	346	366	387	408	2	4	6	8	10	13	15	17	19
211	428	449	469	490	510	531	552	572	593	613	2	4	6	8	10	13	15	17	19
212	634	654	675	695	715	736	756	777	797	818	2	4	6	8	10	12	14	16	18
213	838	858	879	899	919	940	960	980	33001	33021	2	4	6	8	10	12	14	16	18
214	33041	33062	33082	33102	33122	33143	33163	33183	203	224	2	4	6	8	10	12	14	16	18
215	244	264	284	304	325	345	365	385	405	425	2	4	6	8	10	12	14	16	18
216	445	465	486	506	526	546	566	586	606	626	2	4	6	8	10	12	14	16	18
217	646	666	686	706	726	746	766	786	806	826	2	4	6	8	10	12	14	16	18
218	346	866	885	905	925	945	965	985	34005	34025	2	4	6	8	10	12	14	16	18
219	34044	34064	34084	34104	34124	34143	34163	34183	203	223	2	4	6	8	10	12	14	16	18
220	242	262	282	301	321	341	361	380	400	420	2	4	6	8	10	12	14	16	18
221	439	459	479	498	518	537	557	577	596	616	2	4	6	8	10	12	14	16	18
222	635	655	674	694	713	733	753	772	792	811	2	4	6	8	10	12	14	16	18
223	830	850	869	889	908	928	947	967	986	35005	2	4	6	8	10	11	13	15	17
224	35025	35044	35064	35083	35102	35122	35141	35160	35180	199	2	4	6	8	10	11	13	15	17
225	218	238	257	276	295	315	334	353	372	392	2	4	6	8	10	11	13	15	17
226	411	430	449	468	488	507	526	545	564	583	2	4	6	8	10	11	13	15	17
227	603	622	641	660	679	698	717	736	755	774	2	4	6	8	10	11	13	15	17
228	793	813	832	851	870	889	908	927	946	965	2	4	6	8	10	11	13	15	17
229	984	36003	36021	36040	36059	36078	36097	36116	36135	36154	2	4	6	8	10	11	13	15	17
230	36173	192	211	229	248	267	286	305	324	342	2	4	6	8	10	11	13	15	17
231	361	380	399	418	436	455	474	493	511	530	2	4	6	8	10	11	13	15	17
232	549	568	586	605	624	642	661	680	698	717	2	4	6	8	10	11	13	15	17
233	736	754	773	791	810	829	847	866	884	903	2	4	6	8	10	11	13	15	17
234	922	940	959	977	996	37014	37033	37051	37070	37088	2	4	5	7	9	11	13	14	16
235	37107	37125	37144	37162	37181	199	218	236	254	273	2	4	5	7	9	11	13	14	16
236	291	310	328	346	365	383	401	420	438	457	2	4	5	7	9	11	13	14	16
237	475	493	511	530	548	566	585	603	621	639	2	4	5	7	9	11	13	14	16
238	658	676	694	712	731	749	767	785	803	822	2	4	5	7	9	11	13	14	16
239	840	858	876	894	912	931	949	967	985	38003	2	4	5	7	9	11	13	14	16
240	38021	38039	38057	38075	38093	38112	38130	38148	38166	184	2	4	5	7	9	11	13	14	16
241	202	220	238	256	274	292	310	328	346	364	2	4	5	7	9	11	13	14	16
242	382	399	417	435	453	471	489	507	525	543	2	4	5	7	9	11	13	14	16
243	561	578	596	614	632	650	668	686	703	721	2	4	5	7	9	11	13	14	16
244	793	757	775	792	810	828	846	863	881	899	2	4	5	7	9	11	13	14	16
245	917	934	952	970	987	39005	39023	39041	39058	39076	2	4	5	7	9	11	13	14	16
246	39094	39111	39129	39146	39164	182	199	217	235	252	2	4	5	7	9	11	13	14	16
247	270	287	305	322	340	358	375	393	410	428	2	4	5	7	9	11	13	14	16
248	445	463	480	498	515	533	550	568	585	602	2	3	5	7	8	10	12	14	15
249	620	637	655	672	690	707	724	742	759	777	2	3	5	7	8	10	12	14	15

Five - Place Logarithms of Numbers

250 - 299

N	0	1	2	3	4	5	6	7	8	9	Proportional Parts								
											1	2	3	4	5	6	7	8	9
250	39794	39811	39829	39846	39863	39881	39898	39915	39933	39950	2	3	5	7	8	10	12	14	15
251	967	965	40002	40019	40037	40054	40071	40088	40106	40123	2	3	5	7	8	10	12	14	15
252	40140	40157	175	192	209	226	243	261	278	295	2	3	5	7	8	10	12	14	15
253	312	329	346	364	381	398	415	432	449	466	2	3	5	7	8	10	12	14	15
254	483	500	518	535	552	569	586	603	620	637	2	3	5	7	8	10	12	14	15
255	654	671	688	705	722	739	756	773	790	807	2	3	5	7	8	10	12	14	15
256	824	841	858	875	892	909	926	943	960	976	2	3	5	7	8	10	12	14	15
257	993	41010	41027	41044	41061	41078	41095	41111	41128	41145	2	3	5	7	8	10	12	14	15
258	41162	179	196	212	229	246	263	280	296	313	2	3	5	7	8	10	12	14	15
259	330	347	364	380	397	414	430	447	464	481	2	3	5	7	8	10	12	14	15
260	497	514	531	547	564	581	597	614	631	647	2	3	5	7	8	10	12	14	15
261	664	681	697	714	731	747	764	780	797	814	2	3	5	7	8	10	12	14	15
262	830	847	863	880	896	913	929	946	963	979	2	3	5	7	8	10	12	14	15
263	996	42012	42029	42045	42062	42078	42095	42111	42127	42144	2	3	5	6	8	10	11	13	14
264	42160	177	193	210	226	243	259	275	292	308	2	3	5	6	8	10	11	13	14
265	325	341	357	374	390	406	423	439	455	472	2	3	5	6	8	10	11	13	14
266	488	504	521	537	553	570	586	602	619	635	2	3	5	6	8	10	11	13	14
267	651	667	684	700	716	732	749	765	781	797	2	3	5	6	8	10	11	13	14
268	813	830	846	862	878	894	911	927	943	959	2	3	5	6	8	10	11	13	14
269	975	991	43008	43024	43040	43056	43072	43088	43104	43120	2	3	5	6	8	10	11	13	14
270	43136	43152	169	185	201	217	233	249	265	281	2	3	5	6	8	10	11	13	14
271	297	313	329	345	361	377	393	409	425	441	2	3	5	6	8	10	11	13	14
272	457	473	489	505	521	537	553	569	584	600	2	3	5	6	8	10	11	13	14
273	616	632	648	664	680	696	712	727	743	759	2	3	5	6	8	10	11	13	14
274	775	791	807	823	838	854	870	886	902	917	2	3	5	6	8	10	11	13	14
275	933	949	965	981	996	44012	44028	44044	44059	44075	2	3	5	6	8	10	11	13	14
276	44091	44107	44122	44138	44154	170	185	201	217	232	2	3	5	6	8	10	11	13	14
277	248	264	279	295	311	326	342	358	373	389	2	3	5	6	8	10	11	13	14
278	404	420	436	451	467	483	498	514	529	545	2	3	5	6	8	10	11	13	14
279	560	576	592	607	623	638	654	669	685	700	2	3	5	6	8	10	11	13	14
280	716	731	747	762	778	793	809	824	840	855	2	3	4	6	8	9	10	12	14
281	871	886	902	917	932	948	963	979	994	45010	2	3	4	6	8	9	10	12	14
282	45025	45040	45056	45071	45086	45102	45117	45133	45148	163	2	3	4	6	8	9	10	12	14
283	179	194	209	225	240	255	271	286	301	317	2	3	4	6	8	9	10	12	14
284	332	347	362	378	393	408	423	439	454	469	2	3	4	6	8	9	10	12	14
285	484	500	515	530	545	561	576	591	606	621	2	3	4	6	8	9	10	12	14
286	637	652	667	682	697	712	728	743	758	773	2	3	4	6	8	9	10	12	14
287	788	803	818	834	849	864	879	894	909	924	2	3	4	6	8	9	10	12	14
288	939	954	969	984	46000	46015	46030	46045	46060	46075	2	3	4	6	8	9	10	12	14
289	46090	46105	46120	46135	150	165	180	195	210	225	2	3	4	6	8	9	10	12	14
290	240	255	270	285	300	315	330	345	359	374	2	3	4	6	8	9	10	12	14
291	389	404	419	434	449	464	479	494	509	523	2	3	4	6	8	9	10	12	14
292	538	553	568	583	598	613	627	642	657	672	2	3	4	6	8	9	10	12	14
293	687	702	716	731	746	761	776	790	805	820	2	3	4	6	8	9	10	12	14
294	835	850	864	879	894	909	923	938	953	967	2	3	4	6	8	9	10	12	14
295	982	997	47012	47026	47041	47056	47070	47085	47100	47114	2	3	4	6	8	9	10	12	14
296	47129	47144	159	173	188	202	217	232	246	261	2	3	4	6	8	9	10	12	14
297	276	290	305	319	334	349	363	378	392	407	2	3	4	6	8	9	10	12	14
298	422	436	451	465	480	494	509	524	538	553	2	3	4	6	8	9	10	12	14
299	567	582	596	611	625	640	654	669	683	698	2	3	4	6	8	9	10	12	14

Five-Place Logarithms of Numbers

300 - 349

N	0	1	2	3	4	5	6	7	8	9	Proportional Parts									
											1	2	3	4	5	6	7	8	9	
300	47712	47727	47741	47756	47770	47784	47799	47813	47828	47842	1	3	4	6	7	8	10	11	13	
301	857	871	885	900	914	929	943	958	972	986	1	3	4	6	7	8	10	11	13	
302	48001	48015	48029	48044	48058	48073	48087	48101	48116	48130	1	3	4	6	7	8	10	11	13	
303	144	159	173	187	202	216	230	244	259	273	1	3	4	6	7	8	10	11	13	
304	287	302	316	330	344	359	373	387	401	416	1	3	4	6	7	8	10	11	13	
305	430	444	458	473	487	501	515	530	544	558	1	3	4	6	7	8	10	11	13	
306	572	586	601	615	629	642	657	671	686	700	1	3	4	6	7	8	10	11	13	
307	714	728	742	756	770	785	799	813	827	841	1	3	4	6	7	8	10	11	13	
308	855	869	883	897	911	926	940	954	968	982	1	3	4	6	7	8	10	11	13	
309	996	49010	49024	49038	49052	49066	49080	49094	49108	49122	1	3	4	6	7	8	10	11	13	
310	49136	150	164	178	192	206	220	234	248	262	1	3	4	6	7	8	10	11	13	
311	276	290	304	318	332	346	360	374	388	402	1	3	4	6	7	8	10	11	13	
312	415	429	443	457	471	485	499	513	527	541	1	3	4	6	7	8	10	11	13	
313	554	568	582	596	610	624	638	651	665	679	1	3	4	6	7	8	10	11	13	
314	693	707	721	734	748	762	776	790	803	817	1	3	4	6	7	8	10	11	13	
315	831	845	859	872	886	900	914	927	941	955	1	3	4	6	7	8	10	11	13	
316	969	982	996	50010	50024	50037	50051	50065	50079	50092	1	3	4	6	7	8	10	11	13	
317	50106	50120	50133	147	161	174	188	202	215	229	1	3	4	6	7	8	10	11	13	
318	243	256	270	284	297	311	325	338	352	365	1	3	4	6	7	8	10	11	13	
319	379	393	406	420	433	447	461	474	488	501	1	3	4	6	7	8	10	11	13	
320	515	529	542	556	569	583	596	610	623	637	1	3	4	6	7	8	10	11	13	
321	651	664	678	691	705	718	732	745	759	772	1	3	4	6	7	8	10	11	13	
322	786	799	813	826	840	853	866	880	893	907	1	3	4	5	6	8	9	10	12	
323	920	934	947	961	974	987	51001	51014	51028	51041	1	3	4	5	6	8	9	10	12	
324	51055	51068	51081	51095	51108	51121	135	148	162	175	1	3	4	5	6	8	9	10	12	
325	188	202	215	228	242	255	268	282	295	308	1	3	4	5	6	8	9	10	12	
326	322	335	348	362	375	388	402	415	428	441	1	3	4	5	6	8	9	10	12	
327	455	468	481	495	508	521	534	548	561	574	1	3	4	5	6	8	9	10	12	
328	587	601	614	627	640	654	667	680	693	706	1	3	4	5	6	8	9	10	12	
329	720	733	746	759	772	786	799	812	825	838	1	3	4	5	6	8	9	10	12	
330	851	865	878	891	904	917	930	943	957	970	1	3	4	5	6	8	9	10	12	
331	983	996	52009	52022	52035	52048	52061	52075	52088	52101	1	3	4	5	6	8	9	10	12	
332	52114	52127	140	153	166	179	192	205	218	231	1	3	4	5	6	8	9	10	12	
333	244	257	270	284	297	310	323	336	349	362	1	3	4	5	6	8	9	10	12	
334	375	388	401	414	427	440	453	466	479	492	1	3	4	5	6	8	9	10	12	
335	504	517	530	543	556	569	582	595	608	621	1	3	4	5	6	8	9	10	12	
336	634	647	660	673	686	699	711	724	737	750	1	3	4	5	6	8	9	10	12	
337	763	776	789	802	815	827	840	853	866	879	1	3	4	5	6	8	9	10	12	
338	892	905	917	930	943	956	969	982	994	53007	1	3	4	5	6	8	9	10	12	
339	53020	53033	53046	53058	53071	53084	53097	53110	53122	135	1	3	4	5	6	8	9	10	12	
340	148	161	173	186	199	212	224	237	250	263	1	3	4	5	6	8	9	10	12	
341	275	288	301	314	326	339	352	364	377	390	1	3	4	5	6	8	9	10	12	
342	403	415	428	441	453	466	479	491	504	517	1	3	4	5	6	8	9	10	12	
343	529	542	555	567	580	593	605	618	631	643	1	3	4	5	6	8	9	10	12	
344	658	668	681	694	706	719	732	744	757	769	1	3	4	5	6	8	9	10	12	
345	782	794	807	820	832	845	857	870	882	895	1	3	4	5	6	8	9	10	12	
346	908	920	933	945	958	970	983	995	54008	54020	1	2	4	5	6	7	8	10	11	
347	54033	54045	54058	54070	54083	54095	54108	54120	133	145	1	2	4	5	6	7	8	10	11	
348	158	170	183	195	208	220	233	245	258	270	1	2	4	5	6	7	8	10	11	
349	283	295	307	320	332	345	357	370	382	394	1	2	4	5	6	7	8	10	11	

Five - Place Logarithms of Numbers

350 - 399

N	0	1	2	3	4	5	6	7	8	9	Proportional Parts									
											1	2	3	4	5	6	7	8	9	
350	54407	54419	54432	54444	54456	54469	54481	54494	54506	54518	1	2	4	5	6	7	8	10	11	
351	531	543	555	568	580	593	605	617	630	642	1	2	4	5	6	7	8	10	11	
352	654	667	679	691	704	716	728	741	753	765	1	2	4	5	6	7	8	10	11	
353	777	790	802	814	827	839	851	864	876	888	1	2	4	5	6	7	8	10	11	
354	900	913	925	937	949	962	974	986	998	55011	1	2	4	5	6	7	8	10	11	
355	55023	55035	55047	55060	55072	55084	55096	55108	55121	133	1	2	4	5	6	7	8	10	11	
356	145	157	169	182	194	206	218	230	242	255	1	2	4	5	6	7	8	10	11	
357	267	279	291	303	315	328	340	352	364	376	1	2	4	5	6	7	8	10	11	
358	388	400	413	425	437	449	461	473	485	497	1	2	4	5	6	7	8	10	11	
359	509	522	534	546	558	570	582	594	606	618	1	2	4	5	6	7	8	10	11	
360	630	642	654	666	678	691	703	715	727	739	1	2	4	5	6	7	8	10	11	
361	751	763	775	787	799	811	823	835	847	859	1	2	4	5	6	7	8	10	11	
362	871	883	895	907	919	931	943	955	967	979	1	2	4	5	6	7	8	10	11	
363	991	56003	56015	56027	56038	56050	56062	56074	56086	56098	1	2	4	5	6	7	8	10	11	
364	56110	122	134	146	158	170	182	194	205	217	1	2	4	5	6	7	8	10	11	
365	229	241	253	265	277	289	301	312	324	336	1	2	4	5	6	7	8	10	11	
366	348	360	372	384	396	407	419	431	443	455	1	2	4	5	6	7	8	10	11	
367	467	478	490	502	514	526	538	549	561	573	1	2	4	5	6	7	8	10	11	
368	585	597	608	620	632	644	656	667	679	691	1	2	4	5	6	7	8	10	11	
369	703	714	726	738	750	761	773	785	797	808	1	2	4	5	6	7	8	10	11	
370	820	832	844	855	867	879	891	902	914	926	1	2	4	5	6	7	8	10	11	
371	937	949	961	972	984	996	57008	57019	57031	57043	1	2	4	5	6	7	8	10	11	
372	57054	57066	57078	57089	57101	57113	124	136	148	159	1	2	4	5	6	7	8	10	11	
373	171	183	194	206	217	229	241	252	264	276	1	2	4	5	6	7	8	10	11	
374	287	299	310	322	334	345	357	368	380	392	1	2	4	5	6	7	8	10	11	
375	403	415	426	438	449	461	473	484	496	507	1	2	4	5	6	7	8	10	11	
376	519	530	542	553	565	576	588	600	611	623	1	2	4	5	6	7	8	10	11	
377	634	646	657	669	680	692	703	715	726	738	1	2	4	5	6	7	8	10	11	
378	749	761	772	784	795	807	818	830	841	852	1	2	3	4	6	7	8	9	10	
379	864	875	887	898	910	921	933	944	955	967	1	2	3	4	6	7	8	9	10	
380	978	990	58001	58013	58024	58035	58047	58058	58070	58081	1	2	3	4	6	7	8	9	10	
381	58092	58104	115	127	138	149	161	172	184	195	1	2	3	4	6	7	8	9	10	
382	206	218	229	240	252	263	274	286	297	309	1	2	3	4	6	7	8	9	10	
383	320	331	343	354	365	377	388	399	410	422	1	2	3	4	6	7	8	9	10	
384	433	444	456	467	478	490	501	512	524	535	1	2	3	4	6	7	8	9	10	
385	546	557	569	580	591	602	614	625	636	647	1	2	3	4	6	7	8	9	10	
386	659	670	681	692	704	715	726	737	749	760	1	2	3	4	6	7	8	9	10	
387	771	782	794	805	816	827	838	850	861	872	1	2	3	4	6	7	8	9	10	
388	883	894	906	917	928	939	950	961	973	984	1	2	3	4	6	7	8	9	10	
389	995	59006	59017	59028	59040	59051	59062	59073	59084	59095	1	2	3	4	6	7	8	9	10	
390	59016	118	129	140	151	162	173	184	195	207	1	2	3	4	6	7	8	9	10	
391	218	229	240	251	262	273	284	295	306	318	1	2	3	4	6	7	8	9	10	
392	329	340	351	362	373	384	395	406	417	428	1	2	3	4	6	7	8	9	10	
393	439	450	461	472	483	494	506	517	528	539	1	2	3	4	6	7	8	9	10	
394	550	561	572	583	594	605	616	627	638	649	1	2	3	4	6	7	8	9	10	
395	660	671	682	693	704	715	726	737	748	759	1	2	3	4	6	7	8	9	10	
396	770	780	791	802	813	824	835	846	857	868	1	2	3	4	6	7	8	9	10	
397	879	890	901	912	923	934	945	956	966	977	1	2	3	4	6	7	8	9	10	
398	988	999	60010	60021	60032	60043	60054	60065	60076	60086	1	2	3	4	6	7	8	9	10	
399	60097	60108	119	130	141	152	163	173	184	195	1	2	3	4	6	7	8	9	10	

Five - Place Logarithms of Numbers

400 - 449

N	0	1	2	3	4	5	6	7	8	9	Proportional Parts								
											1	2	3	4	5	6	7	8	9
400	60206	60217	60228	60239	60249	60260	60271	60282	60293	60304	1	2	3	4	5	6	7	8	9
401	314	325	336	347	358	369	379	390	401	412	1	2	3	4	5	6	7	8	9
402	423	433	444	455	466	477	487	498	509	520	1	2	3	4	5	6	7	8	9
403	531	541	552	563	574	584	595	606	617	627	1	2	3	4	5	6	7	8	9
404	638	649	660	670	681	692	703	713	724	735	1	2	3	4	5	6	7	8	9
405	746	756	767	778	788	799	810	821	831	842	1	2	3	4	5	6	7	8	9
406	853	863	874	885	895	906	917	927	938	949	1	2	3	4	5	6	7	8	9
407	959	970	981	991	61002	61012	61023	61034	61045	61055	1	2	3	4	5	6	7	8	9
408	61066	61077	61087	61098	109	119	130	140	151	162	1	2	3	4	5	6	7	8	9
409	172	183	194	204	215	225	236	247	257	268	1	2	3	4	5	6	7	8	9
410	278	289	300	310	321	331	342	352	363	374	1	2	3	4	5	6	7	8	9
411	384	395	405	416	426	437	448	458	469	479	1	2	3	4	5	6	7	8	9
412	490	500	511	521	532	542	553	563	574	584	1	2	3	4	5	6	7	8	9
413	595	606	616	627	637	648	658	669	679	690	1	2	3	4	5	6	7	8	9
414	700	711	721	731	742	752	763	773	784	794	1	2	3	4	5	6	7	8	9
415	805	815	826	836	847	857	868	878	888	899	1	2	3	4	5	6	7	8	9
416	909	920	930	941	951	962	972	982	993	62003	1	2	3	4	5	6	7	8	9
417	62014	62024	62034	62045	62055	62066	62076	62086	62097	107	1	2	3	4	5	6	7	8	9
418	118	128	138	149	159	170	180	190	201	211	1	2	3	4	5	6	7	8	9
419	221	232	242	252	263	273	284	294	304	315	1	2	3	4	5	6	7	8	9
420	325	335	346	356	366	377	387	397	408	418	1	2	3	4	5	6	7	8	9
421	428	439	449	459	469	480	490	500	511	521	1	2	3	4	5	6	7	8	9
422	531	542	552	562	572	583	593	603	613	624	1	2	3	4	5	6	7	8	9
423	634	644	655	665	675	685	696	706	716	726	1	2	3	4	5	6	7	8	9
424	737	747	757	767	778	788	798	808	818	829	1	2	3	4	5	6	7	8	9
425	839	849	859	870	880	890	900	910	921	931	1	2	3	4	5	6	7	8	9
426	941	951	961	972	982	992	63002	63012	63022	63033	1	2	3	4	5	6	7	8	9
427	63043	63053	63063	63073	63083	63094	104	114	124	134	1	2	3	4	5	6	7	8	9
428	144	155	165	175	185	195	205	215	225	236	1	2	3	4	5	6	7	8	9
429	246	256	266	276	286	296	306	317	327	337	1	2	3	4	5	6	7	8	9
430	347	357	367	377	387	397	407	417	428	438	1	2	3	4	5	6	7	8	9
431	448	458	468	478	488	498	508	518	528	538	1	2	3	4	5	6	7	8	9
432	548	558	568	579	589	599	609	619	629	639	1	2	3	4	5	6	7	8	9
433	649	659	669	679	689	699	709	719	729	739	1	2	3	4	5	6	7	8	9
434	749	759	769	779	789	799	809	819	829	839	1	2	3	4	5	6	7	8	9
435	849	859	869	879	889	899	909	919	929	939	1	2	3	4	5	6	7	8	9
436	949	959	969	979	988	998	64008	64018	64028	64038	1	2	3	4	5	6	7	8	9
437	64048	64058	64068	64078	64088	64098	108	118	128	137	1	2	3	4	5	6	7	8	9
438	147	157	167	177	187	197	207	217	227	237	1	2	3	4	5	6	7	8	9
439	246	256	266	276	286	296	306	316	326	335	1	2	3	4	5	6	7	8	9
440	345	355	365	375	385	395	404	414	424	434	1	2	3	4	5	6	7	8	9
441	444	454	464	473	483	493	503	513	523	532	1	2	3	4	5	6	7	8	9
442	542	552	562	572	582	591	601	611	621	631	1	2	3	4	5	6	7	8	9
443	640	650	660	670	680	689	699	709	719	729	1	2	3	4	5	6	7	8	9
444	738	748	758	768	777	787	797	807	816	826	1	2	3	4	5	6	7	8	9
445	836	846	856	865	875	885	895	904	914	924	1	2	3	4	5	6	7	8	9
446	933	943	953	963	972	982	992	65002	65011	65021	1	2	3	4	5	6	7	8	9
447	65031	65040	65050	65060	65070	65079	65089	099	108	118	1	2	3	4	5	6	7	8	9
448	128	137	147	157	167	176	186	196	205	215	1	2	3	4	5	6	7	8	9
449	225	234	244	254	263	273	283	292	302	312	1	2	3	4	5	6	7	8	9

Five - Place Logarithms of Numbers

450 - 499

N	0	1	2	3	4	5	6	7	8	9	Proportional Parts								
											1	2	3	4	5	6	7	8	9
450	65321	65331	65341	65350	65360	65369	65379	65389	65398	65408	1	2	3	4	5	6	7	8	9
451	418	427	437	447	456	466	475	485	495	504	1	2	3	4	5	6	7	8	9
452	514	523	533	543	552	562	571	581	591	600	1	2	3	4	5	6	7	8	9
453	610	619	629	639	648	658	667	677	686	696	1	2	3	4	5	6	7	8	9
454	706	715	725	734	744	753	763	772	782	792	1	2	3	4	5	6	7	8	9
455	801	811	820	830	839	849	858	868	877	887	1	2	3	4	5	6	7	8	9
456	896	906	916	925	935	944	954	963	973	982	1	2	3	4	5	6	7	8	9
457	992	66001	66011	66020	66030	66039	66049	66058	66068	66077	1	2	3	4	5	6	7	8	9
458	66087	096	106	115	124	134	143	153	162	172	1	2	3	4	5	6	7	8	9
459	181	191	200	210	219	229	238	247	257	266	1	2	3	4	5	6	7	8	9
460	276	285	295	304	314	323	332	342	351	361	1	2	3	4	5	6	7	8	9
461	370	380	389	398	408	417	427	436	445	455	1	2	3	4	5	6	7	8	9
462	464	474	483	492	502	511	521	530	539	549	1	2	3	4	5	6	7	8	9
463	558	567	577	586	596	605	614	624	633	642	1	2	3	4	5	6	7	8	9
464	652	661	671	680	689	699	708	717	727	736	1	2	3	4	5	6	7	8	9
465	745	755	764	773	783	792	801	811	820	829	1	2	3	4	5	6	7	8	9
466	839	848	857	867	876	885	894	904	913	922	1	2	3	4	5	6	7	8	9
467	932	941	950	960	969	978	987	997	67006	67015	1	2	3	4	5	6	7	8	9
468	67025	67034	67043	67052	67062	67071	67080	67089	099	108	1	2	3	4	5	6	7	8	9
469	117	127	136	145	154	164	173	182	191	201	1	2	3	4	5	6	7	8	9
470	210	219	228	237	247	256	265	274	284	293	1	2	3	4	5	6	7	8	9
471	302	311	321	330	339	348	357	367	376	385	1	2	3	4	5	6	7	8	9
472	394	403	413	422	431	440	449	459	468	477	1	2	3	4	5	6	7	8	9
473	486	495	504	514	523	532	541	550	560	569	1	2	3	4	5	6	7	8	9
474	578	587	596	605	614	624	633	642	651	660	1	2	3	4	5	6	7	8	9
475	669	679	688	697	706	715	724	733	742	752	1	2	3	4	5	6	7	8	9
476	761	770	779	788	797	806	815	825	834	843	1	2	3	4	5	6	7	8	9
477	852	861	870	879	888	897	906	916	925	934	1	2	3	4	5	6	7	8	9
478	943	952	961	970	979	988	997	68006	68015	68024	1	2	3	4	5	6	7	8	9
479	68034	68043	68052	68061	68070	68079	68088	097	106	115	1	2	3	4	5	6	7	8	9
480	124	133	142	151	160	169	178	187	196	205	1	2	3	4	5	6	7	8	9
481	215	224	233	242	251	260	269	278	287	296	1	2	3	4	5	6	7	8	9
482	305	314	323	332	341	350	359	368	377	386	1	2	3	4	5	6	7	8	9
483	395	404	413	422	431	440	449	458	467	476	1	2	3	4	5	6	7	8	9
484	485	494	502	511	520	529	538	547	556	565	1	2	3	4	5	6	7	8	9
485	574	583	592	601	610	619	628	637	646	655	1	2	3	4	5	6	7	8	9
486	664	673	681	690	699	708	717	726	735	744	1	2	3	4	5	6	7	8	9
487	753	762	771	780	789	797	806	815	824	833	1	2	3	4	5	6	7	8	9
488	842	851	860	869	878	886	895	904	913	922	1	2	3	4	5	6	7	8	9
489	931	940	949	958	966	975	984	993	69002	69011	1	2	3	4	5	6	7	8	9
490	69020	69028	69037	69046	69055	69064	69073	69082	090	099	1	2	3	4	5	6	7	8	9
491	108	117	126	135	144	152	161	170	179	188	1	2	3	4	5	6	7	8	9
492	197	205	214	223	232	241	249	258	267	276	1	2	3	4	5	6	7	8	9
493	285	294	302	311	320	329	338	346	355	364	1	2	3	4	5	6	7	8	9
494	373	381	390	399	408	417	425	434	443	452	1	2	3	4	5	6	7	8	9
495	461	469	478	487	496	504	513	522	531	539	1	2	3	4	5	6	7	8	9
496	548	557	566	574	583	592	601	609	618	627	1	2	3	4	5	6	7	8	9
497	636	644	653	662	671	679	688	697	705	714	1	2	3	4	5	6	7	8	9
498	723	732	740	749	758	767	775	784	793	801	1	2	3	4	5	6	7	8	9
499	810	819	827	836	845	854	862	871	880	888	1	2	3	4	5	6	7	8	9

Five - Place Logarithms of Numbers

500 - 549

N	0	1	2	3	4	5	6	7	8	9	Proportional Parts								
											1	2	3	4	5	6	7	8	9
500	69897	69906	69914	69923	69932	69940	69949	69958	69966	69975	1	2	3	4	5	6	7	8	9
501	984	992	70001	70010	70018	70027	70036	70044	70053	70062	1	2	3	4	5	6	7	8	9
502	70070	70079	088	096	105	114	122	131	140	148	1	2	3	4	5	6	7	8	9
503	157	165	174	183	191	200	209	217	226	234	1	2	3	4	5	6	7	8	9
504	243	252	260	269	278	286	295	303	312	321	1	2	3	4	5	6	7	8	9
505	329	338	346	355	364	372	381	389	398	406	1	2	3	4	5	6	7	8	9
506	415	424	432	441	449	458	467	475	484	492	1	2	3	4	5	6	7	8	9
507	501	509	518	526	535	544	552	561	569	578	1	2	3	4	5	6	7	8	9
508	586	595	603	612	621	629	638	646	655	663	1	2	3	4	5	6	7	8	9
509	672	680	689	697	706	714	723	731	740	749	1	2	3	4	5	6	7	8	9
510	757	766	774	783	791	800	808	817	825	834	1	2	3	4	5	6	7	8	9
511	842	851	859	868	876	885	893	902	910	919	1	2	3	4	5	6	7	8	9
512	927	935	944	952	961	969	978	986	995	71003	1	2	2	3	4	5	6	6	7
513	71012	71020	71029	71037	71046	71054	71063	71071	71079	088	1	2	2	3	4	5	6	6	7
514	096	105	113	122	130	139	147	155	164	172	1	2	2	3	4	5	6	6	7
515	181	189	198	206	214	223	231	240	248	257	1	2	2	3	4	5	6	6	7
516	265	273	282	290	299	307	315	324	332	341	1	2	2	3	4	5	6	6	7
517	349	357	366	374	383	391	399	408	416	425	1	2	2	3	4	5	6	6	7
518	433	441	450	458	466	475	483	492	500	508	1	2	2	3	4	5	6	6	7
519	517	525	533	542	550	559	567	575	584	592	1	2	2	3	4	5	6	6	7
520	600	609	617	625	634	642	650	659	667	675	1	2	2	3	4	5	6	6	7
521	684	692	700	709	717	725	734	742	750	759	1	2	2	3	4	5	6	6	7
522	767	775	784	792	800	809	817	825	834	842	1	2	2	3	4	5	6	6	7
523	850	858	867	875	883	892	900	908	917	925	1	2	2	3	4	5	6	6	7
524	933	941	950	958	966	975	983	991	999	72008	1	2	2	3	4	5	6	6	7
525	72016	72024	72032	72041	72049	72057	72066	72074	72082	090	1	2	2	3	4	5	6	6	7
526	099	107	115	123	132	140	149	156	165	173	1	2	2	3	4	5	6	6	7
527	181	189	198	206	214	222	230	239	247	255	1	2	2	3	4	5	6	6	7
528	263	272	280	288	296	304	313	321	329	337	1	2	2	3	4	5	6	6	7
529	346	354	362	370	378	387	395	403	411	419	1	2	2	3	4	5	6	6	7
530	428	436	444	452	460	469	477	485	493	501	1	2	2	3	4	5	6	6	7
531	509	518	526	534	542	550	558	567	575	583	1	2	2	3	4	5	6	6	7
532	591	599	607	616	624	632	640	648	656	665	1	2	2	3	4	5	6	6	7
533	673	681	689	697	705	713	722	730	738	746	1	2	2	3	4	5	6	6	7
534	754	762	770	779	787	795	803	811	819	827	1	2	2	3	4	5	6	6	7
535	835	843	852	860	868	876	884	892	900	908	1	2	2	3	4	5	6	6	7
536	916	925	933	941	949	957	965	973	981	989	1	2	2	3	4	5	6	6	7
537	997	73006	73014	73022	73030	73038	73046	73054	73062	73070	1	2	2	3	4	5	6	6	7
538	73078	086	094	102	111	119	127	135	143	151	1	2	2	3	4	5	6	6	7
539	159	167	175	183	191	199	207	215	223	231	1	2	2	3	4	5	6	6	7
540	239	247	255	263	272	280	288	296	304	312	1	2	2	3	4	5	6	6	7
541	320	328	336	344	352	360	368	376	384	392	1	2	2	3	4	5	6	6	7
542	400	408	416	424	432	440	448	456	464	472	1	2	2	3	4	5	6	6	7
543	480	488	496	504	512	520	528	536	544	552	1	2	2	3	4	5	6	6	7
544	560	568	576	584	592	600	608	616	624	632	1	2	2	3	4	5	6	6	7
545	640	648	656	664	672	679	687	695	703	711	1	2	2	3	4	5	6	6	7
546	719	727	735	743	751	759	767	775	783	791	1	2	2	3	4	5	6	6	7
547	799	807	815	823	830	838	846	854	862	870	1	2	2	3	4	5	6	6	7
548	878	886	894	902	910	918	926	933	941	949	1	2	2	3	4	5	6	6	7
549	957	965	973	981	989	997	74005	74013	74020	74028	1	2	2	3	4	5	6	6	7

Five - Place Logarithms of Numbers

550 - 599

N	0	1	2	3	4	5	6	7	8	9	Proportional Parts								
											1	2	3	4	5	6	7	8	9
550	74036	74044	74052	74060	74068	74076	74084	74092	74099	74107	1	2	2	3	4	5	6	6	7
551	115	123	131	139	147	155	162	170	178	186	1	2	2	3	4	5	6	6	7
552	194	202	210	218	225	233	241	249	257	265	1	2	2	3	4	5	6	6	7
553	273	280	288	296	304	312	320	327	335	343	1	2	2	3	4	5	6	6	7
554	351	359	367	374	382	390	398	406	414	421	1	2	2	3	4	5	6	6	7
555	429	437	445	453	461	468	476	484	492	500	1	2	2	3	4	5	6	6	7
556	507	515	523	531	539	547	554	562	570	578	1	2	2	3	4	5	6	6	7
557	586	593	601	609	617	624	632	640	648	656	1	2	2	3	4	5	6	6	7
558	663	671	679	687	695	702	710	718	726	733	1	2	2	3	4	5	6	6	7
559	741	749	757	764	772	780	788	796	803	811	1	2	2	3	4	5	6	6	7
560	819	827	834	842	850	858	865	873	881	889	1	2	2	3	4	5	6	6	7
561	896	904	912	920	927	935	943	950	958	966	1	2	2	3	4	5	6	6	7
562	974	981	989	997	75005	75012	75020	75028	75035	75043	1	2	2	3	4	5	6	6	7
563	75051	75059	75066	75074	082	089	097	105	113	120	1	2	2	3	4	5	6	6	7
564	128	136	143	151	159	166	174	182	189	197	1	2	2	3	4	5	6	6	7
565	205	213	220	228	236	243	251	259	266	274	1	2	2	3	4	5	6	6	7
566	282	289	297	305	312	320	328	335	343	351	1	2	2	3	4	5	6	6	7
567	358	366	374	381	389	397	404	412	420	427	1	2	2	3	4	5	6	6	7
568	435	442	450	458	465	473	481	488	496	504	1	2	2	3	4	5	6	6	7
569	511	519	526	534	542	549	557	565	572	580	1	2	2	3	4	5	6	6	7
570	587	595	603	610	618	626	633	641	648	656	1	2	2	3	4	5	6	6	7
571	664	671	679	686	694	702	709	717	724	732	1	2	2	3	4	5	6	6	7
572	740	747	755	762	770	778	785	793	800	808	1	2	2	3	4	5	6	6	7
573	815	823	831	838	846	853	861	868	876	884	1	2	2	3	4	5	6	6	7
574	891	899	906	914	921	929	937	944	952	959	1	2	2	3	4	5	6	6	7
575	967	974	982	989	997	76005	76012	76020	76027	76035	1	2	2	3	4	5	6	6	7
576	76042	76050	76057	76065	76072	080	087	095	103	110	1	2	2	3	4	5	6	6	7
577	118	125	133	140	148	155	163	170	178	185	1	1	2	3	4	4	5	6	6
578	193	200	208	215	223	230	238	245	253	260	1	1	2	3	4	4	5	6	6
579	269	275	283	290	298	305	313	320	328	335	1	1	2	3	4	4	5	6	6
580	343	350	358	365	373	380	388	395	403	410	1	1	2	3	4	4	5	6	6
581	418	425	433	440	448	455	462	470	477	485	1	1	2	3	4	4	5	6	6
582	492	500	507	515	522	530	537	545	552	559	1	1	2	3	4	4	5	6	6
583	567	574	582	589	597	604	612	619	626	634	1	1	2	3	4	4	5	6	6
584	641	649	656	664	671	678	686	693	701	708	1	1	2	3	4	4	5	6	6
585	716	723	730	738	745	753	760	768	775	782	1	1	2	3	4	4	5	6	6
586	790	797	805	812	819	827	834	842	849	856	1	1	2	3	4	4	5	6	6
587	864	871	879	886	893	901	908	916	923	930	1	1	2	3	4	4	5	6	6
588	938	945	953	960	967	975	982	989	997	70004	1	1	2	3	4	4	5	6	6
589	77012	77019	77026	77034	77041	77048	77056	77063	77070	078	1	1	2	3	4	4	5	6	6
590	085	093	100	107	115	122	129	137	144	151	1	1	2	3	4	4	5	6	6
591	159	166	173	181	188	195	203	210	217	225	1	1	2	3	4	4	5	6	6
592	232	240	247	254	262	269	276	283	291	298	1	1	2	3	4	4	5	6	6
593	305	313	320	327	335	342	349	357	364	371	1	1	2	3	4	4	5	6	6
594	379	386	393	401	408	415	422	430	437	444	1	1	2	3	4	4	5	6	6
595	452	459	466	474	481	488	495	503	510	517	1	1	2	3	4	4	5	6	6
596	525	532	539	546	554	561	568	576	583	590	1	1	2	3	4	4	5	6	6
597	597	605	612	619	627	634	641	648	656	663	1	1	2	3	4	4	5	6	6
598	670	677	685	692	699	706	714	721	728	735	1	1	2	3	4	4	5	6	6
599	743	750	757	764	772	779	786	793	801	808	1	1	2	3	4	4	5	6	6

Five - Place Logarithms of Numbers

600 - 649

N	0	1	2	3	4	5	6	7	8	9	Proportional Parts								
											1	2	3	4	5	6	7	8	9
600	77815	77822	77830	77837	77844	77851	77859	77866	77873	77880	1	1	2	3	4	4	5	6	6
601	887	895	902	909	916	924	931	938	945	952	1	1	2	3	4	4	5	6	6
602	960	967	974	981	988	996	78003	78010	78017	78025	1	1	2	3	4	4	5	6	6
603	78032	78039	78046	78053	78061	78068	075	082	089	097	1	1	2	3	4	4	5	6	6
604	104	111	118	125	132	140	147	154	161	168	1	1	2	3	4	4	5	6	6
605	176	183	190	197	204	211	219	226	233	240	1	1	2	3	4	4	5	6	6
606	247	254	262	269	276	283	290	297	305	312	1	1	2	3	4	4	5	6	6
607	319	326	333	340	347	355	362	369	376	383	1	1	2	3	4	4	5	6	6
608	390	398	405	412	419	426	433	440	447	455	1	1	2	3	4	4	5	6	6
609	462	469	476	483	490	497	504	512	519	526	1	1	2	3	4	4	5	6	6
610	533	540	547	554	561	569	576	583	590	597	1	1	2	3	4	4	5	6	6
611	604	611	618	625	633	640	647	654	661	668	1	1	2	3	4	4	5	6	6
612	675	682	689	696	704	711	718	725	732	739	1	1	2	3	4	4	5	6	6
613	746	753	760	767	774	781	789	796	803	810	1	1	2	3	4	4	5	6	6
614	817	824	831	838	845	852	859	866	873	880	1	1	2	3	4	4	5	6	6
615	888	895	902	909	916	923	930	937	944	951	1	1	2	3	4	4	5	6	6
616	958	965	972	979	986	993	79000	79007	79014	79021	1	1	2	3	4	4	5	6	6
617	79029	79036	79043	79050	79057	79064	071	078	085	092	1	1	2	3	4	4	5	6	6
618	099	106	113	120	127	134	141	148	155	162	1	1	2	3	4	4	5	6	6
619	169	176	183	190	197	204	211	218	225	232	1	1	2	3	4	4	5	6	6
620	239	246	253	260	267	274	281	288	295	302	1	1	2	3	4	4	5	6	6
621	309	316	323	330	337	344	351	358	365	372	1	1	2	3	4	4	5	6	6
622	379	386	393	400	407	414	421	428	435	442	1	1	2	3	4	4	5	6	6
623	449	456	463	470	477	484	491	498	505	511	1	1	2	3	4	4	5	6	6
624	518	525	532	539	546	553	560	567	574	581	1	1	2	3	4	4	5	6	6
625	588	595	602	609	616	623	630	637	644	650	1	1	2	3	4	4	5	6	6
626	657	664	671	678	685	692	699	706	713	720	1	1	2	3	4	4	5	6	6
627	727	734	741	748	754	761	768	775	782	789	1	1	2	3	4	4	5	6	6
628	796	803	810	817	824	831	837	844	851	858	1	1	2	3	4	4	5	6	6
629	865	872	879	886	893	900	906	913	920	927	1	1	2	3	4	4	5	6	6
630	934	941	948	955	962	969	975	982	989	996	1	1	2	3	4	4	5	6	6
631	80003	80010	80017	80024	80030	80037	80044	80051	80058	80065	1	1	2	3	4	4	5	6	6
632	072	079	085	092	099	106	113	120	127	134	1	1	2	3	4	4	5	6	6
633	140	147	154	161	168	175	182	188	195	202	1	1	2	3	4	4	5	6	6
634	209	216	223	229	236	243	250	257	264	271	1	1	2	3	4	4	5	6	6
635	277	284	291	298	305	312	318	325	332	339	1	1	2	3	4	4	5	6	6
636	346	353	359	366	373	380	387	393	400	407	1	1	2	3	4	4	5	6	6
637	414	421	428	434	441	448	455	462	468	475	1	1	2	3	4	4	5	6	6
638	482	489	496	502	509	516	523	530	536	543	1	1	2	3	4	4	5	6	6
639	550	557	564	570	577	584	591	598	604	611	1	1	2	3	4	4	5	6	6
640	618	625	632	638	645	652	659	665	672	679	1	1	2	3	4	4	5	6	6
641	686	693	699	706	713	720	726	733	740	747	1	1	2	3	4	4	5	6	6
642	754	760	767	774	781	787	794	801	808	814	1	1	2	3	4	4	5	6	6
643	821	828	835	841	848	855	862	868	875	882	1	1	2	3	4	4	5	6	6
644	889	895	902	909	916	922	929	936	943	949	1	1	2	3	4	4	5	6	6
645	956	963	969	976	983	990	996	81003	81010	81017	1	1	2	3	4	4	5	6	6
646	81023	81030	81037	81043	81050	81057	81064	070	077	084	1	1	2	3	4	4	5	6	6
647	090	097	104	111	117	124	131	137	144	151	1	1	2	3	4	4	5	6	6
648	158	164	171	178	184	191	198	204	211	218	1	1	2	3	4	4	5	6	6
649	224	231	238	245	251	258	265	271	278	285	1	1	2	3	4	4	5	6	6

Five-Place Logarithms of Numbers

650 - 699

N											Proportional Parts								
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
650	81291	81298	81305	81311	81318	81325	81331	81338	81345	81351	1	1	2	3	4	4	5	6	6
651	358	365	371	378	385	391	398	405	411	418	1	1	2	3	4	4	5	6	6
652	425	431	438	445	451	458	465	471	478	485	1	1	2	3	4	4	5	6	6
653	491	498	505	511	518	525	531	538	544	551	1	1	2	3	4	4	5	6	6
654	558	564	571	578	584	591	598	604	611	617	1	1	2	3	4	4	5	6	6
655	624	631	637	644	651	657	664	671	677	684	1	1	2	3	4	4	5	6	6
656	690	697	704	710	717	723	730	737	743	750	1	1	2	3	4	4	5	6	6
657	757	763	770	776	783	790	796	803	809	816	1	1	2	3	4	4	5	6	6
658	823	829	836	842	849	856	862	869	875	882	1	1	2	3	4	4	5	6	6
659	889	895	902	908	915	921	928	935	941	948	1	1	2	3	4	4	5	6	6
660	954	961	968	974	981	987	994	82000	82007	82014	1	1	2	3	4	4	5	6	6
661	82020	82027	82033	82040	82046	82053	82060	066	073	079	1	1	2	3	4	4	5	6	6
662	086	092	099	105	112	119	125	132	138	145	1	1	2	3	4	4	5	6	6
663	151	158	164	171	178	184	191	197	204	210	1	1	2	3	4	4	5	6	6
664	217	223	230	236	243	249	256	263	269	276	1	1	2	3	4	4	5	6	6
665	282	289	295	302	308	315	321	328	334	341	1	1	2	3	4	4	5	6	6
666	347	354	360	367	373	380	387	393	400	406	1	1	2	3	4	4	5	6	6
667	413	419	426	432	439	445	452	458	465	471	1	1	2	2	3	4	4	5	5
668	478	484	491	497	504	510	517	523	530	536	1	1	2	2	3	4	4	5	5
669	543	549	556	562	569	575	582	588	595	601	1	1	2	2	3	4	4	5	5
670	607	614	620	627	633	640	646	653	659	666	1	1	2	2	3	4	4	5	5
671	672	679	685	692	698	705	711	718	724	730	1	1	2	2	3	4	4	5	5
672	737	743	750	756	763	769	776	782	789	795	1	1	2	2	3	4	4	5	5
673	802	808	814	821	827	834	840	847	853	860	1	1	2	2	3	4	4	5	5
674	866	872	879	885	892	898	905	911	918	924	1	1	2	2	3	4	4	5	5
675	930	937	943	950	956	962	969	975	982	988	1	1	2	2	3	4	4	5	5
676	995	83001	83008	83014	83020	83027	83033	83040	83046	83052	1	1	2	2	3	4	4	5	5
677	83059	065	072	078	085	091	097	104	110	117	1	1	2	2	3	4	4	5	5
678	123	129	136	142	149	155	161	168	174	181	1	1	2	2	3	4	4	5	5
679	187	193	200	206	213	219	225	232	238	245	1	1	2	2	3	4	4	5	5
680	251	257	264	270	276	283	289	296	302	308	1	1	2	2	3	4	4	5	5
681	315	321	327	334	340	347	353	359	366	372	1	1	2	2	3	4	4	5	5
682	378	385	391	398	404	410	417	423	429	436	1	1	2	2	3	4	4	5	5
683	442	448	455	461	467	474	480	487	493	499	1	1	2	2	3	4	4	5	5
684	506	512	518	525	531	537	544	550	556	563	1	1	2	2	3	4	4	5	5
685	569	575	582	588	594	601	607	613	620	626	1	1	2	2	3	4	4	5	5
686	632	639	645	651	658	664	670	677	683	689	1	1	2	2	3	4	4	5	5
687	696	702	708	715	721	727	734	740	746	753	1	1	2	2	3	4	4	5	5
688	759	765	771	778	784	790	797	803	809	816	1	1	2	2	3	4	4	5	5
689	822	828	835	841	847	853	860	866	872	879	1	1	2	2	3	4	4	5	5
690	885	891	897	904	910	916	923	929	935	942	1	1	2	2	3	4	4	5	5
691	948	954	960	967	973	979	985	992	998	84004	1	1	2	2	3	4	4	5	5
692	84011	84017	84023	84029	84036	84042	84048	84055	84061	067	1	1	2	2	3	4	4	5	5
693	073	080	086	092	098	105	111	117	123	130	1	1	2	2	3	4	4	5	5
694	136	142	148	155	161	167	173	180	186	192	1	1	2	2	3	4	4	5	5
695	198	205	211	217	223	230	236	242	248	255	1	1	2	2	3	4	4	5	5
696	261	267	273	280	286	292	298	305	311	317	1	1	2	2	3	4	4	5	5
697	323	330	336	342	348	354	361	367	373	379	1	1	2	2	3	4	4	5	5
698	386	392	398	404	410	417	423	429	435	442	1	1	2	2	3	4	4	5	5
699	448	454	460	466	473	479	485	491	497	504	1	1	2	2	3	4	4	5	5

Five - Place Logarithms of Numbers

700 - 749

N	0	1	2	3	4	5	6	7	8	9	Proportional Parts								
											1	2	3	4	5	6	7	8	9
700	84510	84516	84522	84528	84535	84541	84547	84553	84559	84566	1	1	2	2	3	4	4	5	5
701	572	578	584	590	597	603	609	615	621	628	1	1	2	2	3	4	4	5	5
702	634	640	646	652	658	665	671	677	683	689	1	1	2	2	3	4	4	5	5
703	696	702	708	714	720	726	733	739	745	751	1	1	2	2	3	4	4	5	5
704	757	763	770	776	782	788	794	800	807	813	1	1	2	2	3	4	4	5	5
705	819	825	831	837	844	850	856	862	868	874	1	1	2	2	3	4	4	5	5
706	880	887	893	899	905	911	917	924	930	936	1	1	2	2	3	4	4	5	5
707	942	948	954	960	967	973	979	985	991	997	1	1	2	2	3	4	4	5	5
708	85003	85009	85016	85022	85028	85034	85040	85046	85052	85058	1	1	2	2	3	4	4	5	5
709	065	071	077	083	089	095	101	107	114	120	1	1	2	2	3	4	4	5	5
710	126	132	138	144	150	156	163	169	175	181	1	1	2	2	3	4	4	5	5
711	187	193	199	205	211	217	224	230	236	242	1	1	2	2	3	4	4	5	5
712	248	254	260	266	272	278	285	291	297	303	1	1	2	2	3	4	4	5	5
713	309	315	321	327	333	339	345	352	358	364	1	1	2	2	3	4	4	5	5
714	370	376	382	388	394	400	406	412	418	425	1	1	2	2	3	4	4	5	5
715	431	437	443	449	455	461	467	473	479	485	1	1	2	2	3	4	4	5	5
716	491	497	503	509	516	522	528	534	540	546	1	1	2	2	3	4	4	5	5
717	552	558	564	570	576	582	588	594	600	606	1	1	2	2	3	4	4	5	5
718	612	618	625	631	637	643	649	655	661	667	1	1	2	2	3	4	4	5	5
719	673	679	685	691	697	703	709	715	721	727	1	1	2	2	3	4	4	5	5
720	733	739	745	751	757	763	769	775	781	788	1	1	2	2	3	4	4	5	5
721	794	800	806	812	818	824	830	836	842	848	1	1	2	2	3	4	4	5	5
722	854	860	866	872	878	884	890	896	902	908	1	1	2	2	3	4	4	5	5
723	914	920	926	932	938	944	950	956	962	968	1	1	2	2	3	4	4	5	5
724	974	980	986	992	998	86004	86010	86016	86022	86028	1	1	2	2	3	4	4	5	5
725	86034	86040	86046	86052	86058	064	070	076	082	088	1	1	2	2	3	4	4	5	5
726	094	100	106	112	118	124	130	136	141	147	1	1	2	2	3	4	4	5	5
727	153	159	165	171	177	183	189	195	201	207	1	1	2	2	3	4	4	5	5
728	213	219	225	231	237	243	249	255	261	267	1	1	2	2	3	4	4	5	5
729	273	279	285	291	297	303	309	314	320	326	1	1	2	2	3	4	4	5	5
730	332	338	344	350	356	362	368	374	380	386	1	1	2	2	3	4	4	5	5
731	392	398	404	410	415	421	427	433	439	445	1	1	2	2	3	4	4	5	5
732	451	457	463	469	475	481	487	493	499	504	1	1	2	2	3	4	4	5	5
733	510	516	522	528	534	540	546	552	558	564	1	1	2	2	3	4	4	5	5
734	570	576	581	587	593	599	605	611	617	623	1	1	2	2	3	4	4	5	5
735	629	635	641	646	652	658	664	670	676	682	1	1	2	2	3	4	4	5	5
736	688	694	700	705	711	717	723	729	735	741	1	1	2	2	3	4	4	5	5
737	747	753	759	764	770	776	782	788	794	800	1	1	2	2	3	4	4	5	5
738	806	812	817	823	829	835	841	847	853	859	1	1	2	2	3	4	4	5	5
739	864	870	876	882	888	894	900	906	911	917	1	1	2	2	3	4	4	5	5
740	923	929	935	941	947	953	958	964	970	976	1	1	2	2	3	4	4	5	5
741	982	988	994	999	87005	87011	87017	87023	87029	87035	1	1	2	2	3	4	4	5	5
742	87040	87046	87052	87058	064	070	075	081	087	093	1	1	2	2	3	4	4	5	5
743	099	105	111	116	122	128	134	140	146	151	1	1	2	2	3	4	4	5	5
744	157	163	169	175	181	186	192	198	204	210	1	1	2	2	3	4	4	5	5
745	216	221	227	233	239	245	251	256	262	268	1	1	2	2	3	4	4	5	5
746	274	280	286	291	297	303	309	315	320	326	1	1	2	2	3	4	4	5	5
747	332	338	344	349	355	361	367	373	379	384	1	1	2	2	3	4	4	5	5
748	390	396	402	408	413	419	425	431	437	442	1	1	2	2	3	4	4	5	5
749	448	454	460	466	471	477	483	489	495	500	1	1	2	2	3	4	4	5	5

Five-Place Logarithms of Numbers

750 - 799

N	0	1	2	3	4	5	6	7	8	9	Proportional Parts								
											1	2	3	4	5	6	7	8	9
750	87506	87512	87518	87523	87529	87535	87541	87547	87552	87558	1	1	2	2	3	4	4	5	5
751	564	570	576	581	587	593	599	604	610	616	1	1	2	2	3	4	4	5	5
752	622	628	633	639	645	651	656	662	668	674	1	1	2	2	3	4	4	5	5
753	679	685	691	697	703	708	714	720	726	731	1	1	2	2	3	4	4	5	5
754	737	743	749	754	760	766	772	777	783	789	1	1	2	2	3	4	4	5	5
755	795	800	806	812	818	823	829	835	841	846	1	1	2	2	3	4	4	5	5
756	852	858	864	869	875	881	887	892	898	904	1	1	2	2	3	4	4	5	5
757	910	915	921	927	933	938	944	950	955	961	1	1	2	2	3	4	4	5	5
758	967	973	978	984	990	996	88001	88007	88013	88018	1	1	2	2	3	4	4	5	5
759	88024	88030	88036	88041	88047	88053	058	064	070	076	1	1	2	2	3	4	4	5	5
760	081	087	093	098	104	110	116	121	127	133	1	1	2	2	3	4	4	5	5
761	138	144	150	156	161	167	173	178	184	190	1	1	2	2	3	4	4	5	5
762	195	201	207	213	218	224	230	235	241	247	1	1	2	2	3	4	4	5	5
763	252	258	264	270	275	281	287	292	298	304	1	1	2	2	3	4	4	5	5
764	309	315	321	326	332	338	343	349	355	360	1	1	2	2	3	4	4	5	5
765	366	372	377	383	389	395	400	406	412	417	1	1	2	2	3	4	4	5	5
766	423	429	434	440	446	451	457	463	468	474	1	1	2	2	3	4	4	5	5
767	480	485	491	497	502	508	513	519	525	530	1	1	2	2	3	4	4	5	5
768	536	542	547	553	559	564	570	576	581	587	1	1	2	2	3	4	4	5	5
769	593	598	604	610	615	621	627	632	638	643	1	1	2	2	3	4	4	5	5
770	649	655	660	666	672	677	683	689	694	700	1	1	2	2	3	4	4	5	5
771	705	711	717	722	728	734	739	745	750	756	1	1	2	2	3	4	4	5	5
772	762	767	773	779	784	790	795	801	807	812	1	1	2	2	3	4	4	5	5
773	818	824	829	835	840	846	852	857	863	868	1	1	2	2	3	4	4	5	5
774	874	880	885	891	897	902	908	913	919	925	1	1	2	2	3	4	4	5	5
775	930	936	941	947	953	958	964	969	975	981	1	1	2	2	3	4	4	5	5
776	986	992	997	89003	89009	89014	89020	89025	89031	89037	1	1	2	2	3	4	4	5	5
777	89042	89048	89053	059	064	070	076	081	087	092	1	1	2	2	3	4	4	5	5
778	098	104	109	115	120	126	131	137	143	148	1	1	2	2	3	4	4	5	5
779	154	159	165	170	176	182	187	193	198	204	1	1	2	2	3	4	4	5	5
780	209	215	221	226	232	237	243	248	254	260	1	1	2	2	3	4	4	5	5
781	265	271	276	282	287	293	298	304	310	315	1	1	2	2	3	4	4	5	5
782	321	326	332	337	343	348	354	360	365	371	1	1	2	2	3	4	4	5	5
783	376	382	387	393	398	404	409	415	421	426	1	1	2	2	3	4	4	5	5
784	432	437	443	448	454	459	465	470	476	481	1	1	2	2	3	4	4	5	5
785	487	492	498	504	509	515	520	526	531	537	1	1	2	2	3	4	4	5	5
786	542	548	553	559	564	570	575	581	586	592	1	1	2	2	3	4	4	5	5
787	597	603	609	614	620	625	631	636	642	647	1	1	2	2	3	4	4	5	5
788	653	658	664	669	675	680	686	691	697	702	0	1	2	2	2	3	4	4	4
789	708	713	719	724	730	735	741	746	752	757	0	1	2	2	2	3	4	4	4
790	763	768	774	779	785	790	796	801	807	812	0	1	2	2	2	3	4	4	4
791	818	823	829	834	840	845	851	856	862	867	0	1	2	2	2	3	4	4	4
792	873	878	883	889	894	900	905	911	916	922	0	1	2	2	2	3	4	4	4
793	927	933	938	944	949	955	960	966	971	977	0	1	2	2	2	3	4	4	4
794	982	988	993	998	90004	90009	90015	90020	90026	90031	0	1	2	2	2	3	4	4	4
795	90037	90042	90048	90053	059	064	069	075	080	086	0	1	2	2	2	3	4	4	4
796	091	097	102	108	113	119	124	129	135	140	0	1	2	2	2	3	4	4	4
797	146	151	157	162	168	173	179	184	189	195	0	1	2	2	2	3	4	4	4
798	200	206	211	217	222	227	233	238	244	249	0	1	2	2	2	3	4	4	4
799	255	260	266	271	276	282	287	293	298	304	0	1	2	2	2	3	4	4	4

Five-Place Logarithms of Numbers

800 - 849

N	0	1	2	3	4	5	6	7	8	9	Proportional Parts								
											1	2	3	4	5	6	7	8	9
800	90309	90314	90320	90325	90331	90336	90342	90347	90352	90358	0	1	2	2	2	3	4	4	4
801	363	369	374	380	385	390	396	401	407	412	0	1	2	2	2	3	4	4	4
802	417	423	428	434	439	445	450	455	461	466	0	1	2	2	2	3	4	4	4
803	472	477	482	488	493	499	504	509	515	520	0	1	2	2	2	3	4	4	4
804	526	531	536	542	547	553	558	563	569	574	0	1	2	2	2	3	4	4	4
805	580	585	590	596	601	607	612	617	623	628	0	1	2	2	2	3	4	4	4
806	634	639	644	650	655	660	666	671	677	682	0	1	2	2	2	3	4	4	4
807	687	693	698	703	709	714	720	725	730	736	0	1	2	2	2	3	4	4	4
808	741	747	752	757	763	768	773	779	784	789	0	1	2	2	2	3	4	4	4
809	795	800	806	811	816	822	827	832	838	843	0	1	2	2	2	3	4	4	4
810	849	854	859	865	870	875	881	886	891	897	0	1	2	2	2	3	4	4	4
811	902	907	913	918	924	929	934	940	945	950	0	1	2	2	2	3	4	4	4
812	956	961	966	972	977	982	988	993	998	91004	0	1	2	2	2	3	4	4	4
813	91009	91014	91020	91025	91030	91036	91041	91046	91052	057	0	1	2	2	2	3	4	4	4
814	062	068	073	078	084	089	094	100	105	110	0	1	2	2	2	3	4	4	4
815	116	121	126	132	137	142	148	153	158	164	0	1	2	2	2	3	4	4	4
816	169	174	180	185	190	196	201	206	212	217	0	1	2	2	2	3	4	4	4
817	222	228	233	238	243	249	254	259	265	270	0	1	2	2	2	3	4	4	4
818	275	281	286	291	297	302	307	312	318	323	0	1	2	2	2	3	4	4	4
819	328	334	339	344	350	355	360	365	371	376	0	1	2	2	2	3	4	4	4
820	381	387	392	397	403	408	413	418	424	429	0	1	2	2	2	3	4	4	4
821	434	440	445	450	455	461	466	471	477	482	0	1	2	2	2	3	4	4	4
822	487	492	498	503	508	514	519	524	529	535	0	1	2	2	2	3	4	4	4
823	540	545	551	556	561	566	572	577	582	587	0	1	2	2	2	3	4	4	4
824	593	598	603	609	614	619	624	630	635	640	0	1	2	2	2	3	4	4	4
825	645	651	656	661	666	672	677	682	687	693	0	1	2	2	2	3	4	4	4
826	698	703	709	714	719	724	730	735	740	745	0	1	2	2	2	3	4	4	4
827	751	756	761	766	772	777	782	787	793	798	0	1	2	2	2	3	4	4	4
828	803	808	814	819	824	829	834	840	845	850	0	1	2	2	2	3	4	4	4
829	855	861	866	871	876	882	887	892	897	903	0	1	2	2	2	3	4	4	4
830	908	913	918	924	929	934	939	944	950	955	0	1	2	2	2	3	4	4	4
831	960	965	971	976	981	986	991	997	92002	92007	0	1	2	2	2	3	4	4	4
832	92012	92018	92023	92028	92033	92038	92044	92049	054	059	0	1	2	2	2	3	4	4	4
833	064	070	075	080	085	091	096	101	106	111	0	1	2	2	2	3	4	4	4
834	117	122	127	132	137	143	148	153	158	163	0	1	2	2	2	3	4	4	4
835	169	174	179	184	189	195	200	205	210	215	0	1	2	2	2	3	4	4	4
836	221	226	231	236	241	247	252	257	262	267	0	1	2	2	2	3	4	4	4
837	273	278	283	288	293	298	304	309	314	319	0	1	2	2	2	3	4	4	4
838	324	330	335	340	345	350	355	361	366	371	0	1	2	2	2	3	4	4	4
839	376	381	387	392	397	402	407	412	418	423	0	1	2	2	2	3	4	4	4
840	428	433	438	443	448	454	459	464	469	474	0	1	2	2	2	3	4	4	4
841	480	485	490	495	500	505	511	516	521	526	0	1	2	2	2	3	4	4	4
842	531	536	542	547	552	557	562	567	572	578	0	1	2	2	2	3	4	4	4
843	583	588	593	598	603	609	614	619	624	629	0	1	2	2	2	3	4	4	4
844	634	639	645	650	655	660	665	670	675	681	0	1	2	2	2	3	4	4	4
845	686	691	696	701	706	711	716	722	727	732	0	1	2	2	2	3	4	4	4
846	737	742	747	752	758	763	768	773	778	783	0	1	2	2	2	3	4	4	4
847	788	793	799	804	809	814	819	824	829	834	0	1	2	2	2	3	4	4	4
848	840	845	850	855	860	865	870	875	881	886	0	1	2	2	2	3	4	4	4
849	891	896	901	906	911	916	921	927	932	937	0	1	2	2	2	3	4	4	4

Five - Place Logarithms of Numbers

850 - 899

N	0	1	2	3	4	5	6	7	8	9	Proportional Parts									
											1	2	3	4	5	6	7	8	9	
850	92942	92947	92952	92957	92962	92967	92973	92978	92983	92988	0	1	2	2	2	3	4	4	4	
851	993	998	93003	93008	93013	93018	93024	93029	93034	93039	0	1	2	2	2	3	4	4	4	
852	93044	93049	054	059	064	069	075	080	085	090	0	1	2	2	2	3	4	4	4	
853	095	100	105	110	115	120	125	131	136	141	0	1	2	2	2	3	4	4	4	
854	146	151	156	161	166	171	176	181	186	192	0	1	2	2	2	3	4	4	4	
855	197	202	207	212	217	222	227	232	237	242	0	1	2	2	2	3	4	4	4	
856	247	252	258	263	268	273	278	283	288	293	0	1	2	2	2	3	4	4	4	
857	298	303	308	313	318	323	328	334	339	344	0	1	2	2	2	3	4	4	4	
858	349	354	359	364	369	374	379	384	389	394	0	1	2	2	2	3	4	4	4	
859	399	404	409	414	420	425	430	435	440	445	0	1	2	2	2	3	4	4	4	
860	450	455	460	465	470	475	480	485	490	495	0	1	2	2	2	3	4	4	4	
861	500	505	510	515	520	526	531	536	541	546	0	1	2	2	2	3	4	4	4	
862	551	556	561	566	571	576	581	586	591	596	0	1	2	2	2	3	4	4	4	
863	601	606	611	616	621	626	631	636	641	646	0	1	2	2	2	3	4	4	4	
864	651	656	661	666	671	676	682	687	692	697	0	1	2	2	2	3	4	4	4	
865	702	707	712	717	722	727	732	737	742	747	0	1	2	2	2	3	4	4	4	
866	752	757	762	767	772	777	782	787	792	797	0	1	2	2	2	3	4	4	4	
867	802	807	812	817	822	827	832	837	842	847	0	1	2	2	2	3	4	4	4	
868	852	857	862	867	872	877	882	887	892	897	0	1	2	2	2	3	4	4	4	
869	902	907	912	917	922	927	932	937	942	947	0	1	2	2	2	3	4	4	4	
870	952	957	962	967	972	977	982	987	992	997	0	1	2	2	2	3	4	4	4	
871	94002	94007	94012	94017	94022	94027	94032	94037	94042	94047	0	1	2	2	2	3	4	4	4	
872	052	057	062	067	072	077	082	086	091	096	0	1	2	2	2	3	4	4	4	
873	101	106	111	116	121	126	131	136	141	146	0	1	2	2	2	3	4	4	4	
874	151	156	161	166	171	176	181	186	191	196	0	1	2	2	2	3	4	4	4	
875	201	206	211	216	221	226	231	236	240	245	0	1	2	2	2	3	4	4	4	
876	250	255	260	265	270	275	280	285	290	295	0	1	2	2	2	3	4	4	4	
877	300	305	310	315	320	325	330	335	340	345	0	1	2	2	2	3	4	4	4	
878	349	354	359	364	369	374	379	384	389	394	0	1	2	2	2	3	4	4	4	
879	399	404	409	414	419	424	429	433	438	443	0	1	2	2	2	3	4	4	4	
880	448	453	458	463	468	473	478	483	488	493	0	1	2	2	2	3	4	4	4	
881	498	503	507	512	517	522	527	532	537	542	0	1	2	2	2	3	4	4	4	
882	547	552	557	562	567	571	576	581	586	591	0	1	2	2	2	3	4	4	4	
883	596	601	606	611	616	621	626	630	635	640	0	1	2	2	2	3	4	4	4	
884	645	650	655	660	665	670	675	680	685	689	0	1	2	2	2	3	4	4	4	
885	694	699	704	709	714	719	724	729	734	738	0	1	2	2	2	3	4	4	4	
886	743	748	753	758	763	768	773	778	783	787	0	1	2	2	2	3	4	4	4	
887	792	797	802	807	812	817	822	827	832	836	0	1	2	2	2	3	4	4	4	
888	841	846	851	856	861	866	871	876	880	885	0	1	2	2	2	3	4	4	4	
889	890	895	900	905	910	915	919	924	929	934	0	1	2	2	2	3	4	4	4	
890	939	944	949	954	959	963	968	973	978	983	0	1	2	2	2	3	4	4	4	
891	988	993	998	95002	95007	95012	95017	95022	95027	95032	0	1	2	2	2	3	4	4	4	
892	95036	95041	95046	051	056	061	066	071	075	080	0	1	2	2	2	3	4	4	4	
893	085	090	095	100	105	109	114	119	124	129	0	1	2	2	2	3	4	4	4	
894	134	139	143	148	153	158	163	168	173	177	0	1	2	2	2	3	4	4	4	
895	182	187	192	197	202	207	211	216	221	226	0	1	2	2	2	3	4	4	4	
896	231	236	240	245	250	255	260	265	270	274	0	1	2	2	2	3	4	4	4	
897	279	284	289	294	299	303	308	313	318	323	0	1	2	2	2	3	4	4	4	
898	328	332	337	342	347	352	357	361	366	371	0	1	2	2	2	3	4	4	4	
899	376	381	386	390	395	400	405	410	415	419	0	1	2	2	2	3	4	4	4	

Five - Place Logarithms of Numbers

900 - 949

N											Proportional Parts									
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	
900	95424	95429	95434	95439	95444	95448	95453	95458	95463	95468	0	1	2	2	2	3	4	4	4	
901	472	477	482	487	492	497	501	506	511	516	0	1	2	2	2	3	4	4	4	
902	521	525	530	535	540	545	550	554	559	564	0	1	2	2	2	3	4	4	4	
903	569	574	578	583	588	593	598	602	607	612	0	1	2	2	2	3	4	4	4	
904	617	622	626	631	636	641	646	650	655	660	0	1	2	2	2	3	4	4	4	
905	665	670	674	679	684	689	694	698	703	708	0	1	2	2	2	3	4	4	4	
906	713	718	722	727	732	737	742	746	751	756	0	1	2	2	2	3	4	4	4	
907	761	766	770	775	780	785	789	794	799	804	0	1	2	2	2	3	4	4	4	
908	809	813	818	823	828	832	837	842	847	852	0	1	2	2	2	3	4	4	4	
909	856	861	866	871	875	880	885	890	895	899	0	1	2	2	2	3	4	4	4	
910	904	909	914	918	923	928	933	938	942	947	0	1	2	2	2	3	4	4	4	
911	952	957	961	966	971	976	980	985	990	995	0	1	2	2	2	3	4	4	4	
912	999	96004	96009	96014	96019	96023	96028	96033	96038	96042	0	1	2	2	2	3	4	4	4	
913	96047	052	057	061	066	071	076	080	085	090	0	1	2	2	2	3	4	4	4	
914	095	099	104	109	114	118	123	128	133	137	0	1	2	2	2	3	4	4	4	
915	142	147	152	156	161	166	171	175	180	185	0	1	2	2	2	3	4	4	4	
916	190	194	199	204	209	213	218	223	227	232	0	1	2	2	2	3	4	4	4	
917	237	242	246	251	256	261	265	270	275	280	0	1	2	2	2	3	4	4	4	
918	284	289	294	298	303	308	313	317	322	327	0	1	2	2	2	3	4	4	4	
919	332	336	341	346	350	355	360	365	369	374	0	1	2	2	2	3	4	4	4	
920	379	384	388	393	398	402	407	412	417	421	0	1	2	2	2	3	4	4	4	
921	426	431	435	440	445	450	454	459	464	468	0	1	2	2	2	3	4	4	4	
922	473	478	483	487	492	497	501	506	511	515	0	1	2	2	2	3	4	4	4	
923	520	525	530	534	539	544	548	553	558	562	0	1	2	2	2	3	4	4	4	
924	567	572	577	581	586	591	595	600	605	609	0	1	2	2	2	3	4	4	4	
925	614	619	624	628	633	638	642	647	652	656	0	1	2	2	2	3	4	4	4	
926	661	666	670	675	680	685	689	694	699	703	0	1	2	2	2	3	4	4	4	
927	708	713	717	722	727	731	736	741	745	750	0	1	2	2	2	3	4	4	4	
928	755	759	764	769	774	778	783	788	792	797	0	1	2	2	2	3	4	4	4	
929	802	806	811	816	820	825	830	834	839	844	0	1	2	2	2	3	4	4	4	
930	848	853	858	862	867	872	876	881	886	890	0	1	2	2	2	3	4	4	4	
931	895	900	904	909	914	918	923	928	932	937	0	1	2	2	2	3	4	4	4	
932	942	946	951	956	960	965	970	974	979	984	0	1	2	2	2	3	4	4	4	
933	988	993	997	97002	97007	97011	97016	97021	97025	97030	0	1	2	2	2	3	4	4	4	
934	97035	97039	97044	049	053	058	063	067	072	077	0	1	2	2	2	3	4	4	4	
935	081	086	090	095	100	104	109	114	118	123	0	1	2	2	2	3	4	4	4	
936	128	132	137	142	146	151	155	160	165	169	0	1	2	2	2	3	4	4	4	
937	174	179	183	188	192	197	202	206	211	216	0	1	2	2	2	3	4	4	4	
938	220	225	230	234	239	243	248	253	257	262	0	1	2	2	2	3	4	4	4	
939	267	271	276	280	285	290	294	299	304	308	0	1	2	2	2	3	4	4	4	
940	313	317	322	327	331	336	340	345	350	354	0	1	2	2	2	3	4	4	4	
941	359	364	368	373	377	382	387	391	396	400	0	1	2	2	2	3	4	4	4	
942	405	410	414	419	424	428	433	437	442	447	0	1	2	2	2	3	4	4	4	
943	451	456	460	465	470	474	479	483	488	493	0	1	2	2	2	3	4	4	4	
944	497	502	506	511	516	520	525	529	534	539	0	1	2	2	2	3	4	4	4	
945	543	548	552	557	562	566	571	575	580	585	0	1	2	2	2	3	4	4	4	
946	589	594	598	603	607	612	617	621	626	630	0	1	2	2	2	3	4	4	4	
947	635	640	644	649	653	658	663	667	672	676	0	1	2	2	2	3	4	4	4	
948	681	685	690	695	699	704	708	713	717	722	0	1	2	2	2	3	4	4	4	
949	727	731	736	740	745	749	754	759	763	768	0	1	2	2	2	3	4	4	4	

Five - Place Logarithms of Numbers

950 - 999

N	0	1	2	3	4	5	6	7	8	9	Proportional Parts								
											1	2	3	4	5	6	7	8	9
950	97772	97777	97782	97786	97791	97795	97800	97804	97809	97813	0	1	2	2	2	3	4	4	4
951	818	823	827	832	836	841	845	850	855	859	0	1	2	2	2	3	4	4	4
952	864	868	873	877	882	886	891	896	900	905	0	1	2	2	2	3	4	4	4
953	909	914	918	923	928	932	937	941	946	950	0	1	2	2	2	3	4	4	4
954	955	959	964	968	973	978	982	987	991	996	0	1	2	2	2	3	4	4	4
955	98000	98005	98009	98014	98019	98023	98028	98032	98037	98041	0	1	2	2	2	3	4	4	4
956	046	050	055	059	064	068	073	078	082	087	0	1	2	2	2	3	4	4	4
957	091	096	100	105	109	114	118	123	127	132	0	1	2	2	2	3	4	4	4
958	137	141	146	150	155	159	164	168	173	177	0	1	2	2	2	3	4	4	4
959	182	186	191	195	200	204	209	214	218	223	0	1	2	2	2	3	4	4	4
960	227	232	236	241	245	250	254	259	263	268	0	1	2	2	2	3	4	4	4
961	272	277	281	286	290	295	299	304	308	313	0	1	2	2	2	3	4	4	4
962	318	322	327	331	336	340	345	349	354	358	0	1	1	2	2	2	3	3	4
963	363	367	372	376	381	385	390	394	399	403	0	1	1	2	2	2	3	3	4
964	408	412	417	421	426	430	435	439	444	448	0	1	1	2	2	2	3	3	4
965	453	457	462	466	471	475	480	484	489	493	0	1	1	2	2	2	3	3	4
966	498	502	507	511	516	520	525	529	534	538	0	1	1	2	2	2	3	3	4
967	543	547	552	556	561	565	570	574	579	583	0	1	1	2	2	2	3	3	4
968	588	592	597	601	605	610	614	619	623	628	0	1	1	2	2	2	3	3	4
969	632	637	641	646	650	655	659	664	668	673	0	1	1	2	2	2	3	3	4
970	677	682	686	691	695	700	704	709	713	717	0	1	1	2	2	2	3	3	4
971	722	726	731	735	740	744	749	753	758	762	0	1	1	2	2	2	3	3	4
972	767	771	776	780	784	789	793	798	802	807	0	1	1	2	2	2	3	3	4
973	811	816	820	825	829	834	838	843	847	851	0	1	1	2	2	2	3	3	4
974	856	860	865	869	874	878	883	887	892	896	0	1	1	2	2	2	3	3	4
975	900	905	909	914	918	923	927	932	936	941	0	1	1	2	2	2	3	3	4
976	945	949	954	958	963	967	972	976	981	985	0	1	1	2	2	2	3	3	4
977	989	994	998	99003	99007	99012	99016	99021	99025	99029	0	1	1	2	2	2	3	3	4
978	99034	99038	99043	047	052	056	061	065	069	074	0	1	1	2	2	2	3	3	4
979	078	083	087	092	096	100	105	109	114	118	0	1	1	2	2	2	3	3	4
980	123	127	131	136	140	145	149	154	158	162	0	1	1	2	2	2	3	3	4
981	167	171	176	180	185	189	193	198	202	207	0	1	1	2	2	2	3	3	4
982	211	216	220	224	229	233	238	242	247	251	0	1	1	2	2	2	3	3	4
983	255	260	264	269	273	277	282	286	291	295	0	1	1	2	2	2	3	3	4
984	300	304	308	313	317	322	326	330	335	339	0	1	1	2	2	2	3	3	4
985	344	348	352	357	361	366	370	374	379	383	0	1	1	2	2	2	3	3	4
986	388	392	396	401	405	410	414	419	423	427	0	1	1	2	2	2	3	3	4
987	432	436	441	445	449	454	458	463	467	471	0	1	1	2	2	2	3	3	4
988	476	480	484	489	493	498	502	506	511	515	0	1	1	2	2	2	3	3	4
989	520	524	528	533	537	542	546	550	555	559	0	1	1	2	2	2	3	3	4
990	564	568	572	577	581	585	590	594	599	603	0	1	1	2	2	2	3	3	4
991	607	612	616	621	625	629	634	638	642	647	0	1	1	2	2	2	3	3	4
992	651	656	660	664	669	673	677	682	686	691	0	1	1	2	2	2	3	3	4
993	695	699	704	708	712	717	721	726	730	734	0	1	1	2	2	2	3	3	4
994	739	744	747	752	756	760	765	769	774	778	0	1	1	2	2	2	3	3	4
995	782	787	791	795	800	804	808	813	817	822	0	1	1	2	2	2	3	3	4
996	826	830	835	839	843	848	852	856	861	865	0	1	1	2	2	2	3	3	4
997	870	874	878	883	887	891	896	900	904	909	0	1	1	2	2	2	3	3	4
998	913	917	922	926	930	935	939	944	948	952	0	1	1	2	2	2	3	3	4
999	957	961	965	970	974	978	983	987	991	996	0	1	1	2	2	2	3	3	4

**SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000¹**

n	n^2	$n^{1/2}$	$1/n$
1	1	1.000 0000	1.000 000 000
2	4	1.414 2136	0.500 000 000
3	9	1.732 0508	.333 333 333
4	16	2.000 0000	.250 000 000
5	25	2.236 0680	.200 000 000
6	36	2.449 4897	.166 666 667
7	49	2.645 7513	.142 857 143
8	64	2.828 4271	.125 000 000
9	81	3.000 0000	.111 111 111
10	1 00	3.162 2777	.100 000 000
11	1 21	3.316 6248	.090 909 091
12	1 44	3.464 1016	.083 333 333
13	1 69	3.605 5513	.076 923 077
14	1 96	3.741 6574	.071 428 571
15	2 25	3.872 9833	.066 666 667
16	2 56	4.000 0000	.062 500 000
17	2 89	4.123 1056	.058 823 529
18	3 24	4.242 6407	.055 555 556
19	3 61	4.358 8989	.052 631 579
20	4 00	4.472 1360	.050 000 000
21	4 41	4.582 5757	.047 619 048
22	4 84	4.690 4158	.045 454 545
23	5 29	4.795 8315	.043 478 261
24	5 76	4.898 9795	.041 666 667
25	6 25	5.000 0000	.040 000 000
26	6 76	5.099 0195	.038 461 538
27	7 29	5.196 1524	.037 037 037
28	7 84	5.291 5026	.035 714 286
29	8 41	5.385 1648	.034 482 759
30	9 00	5.477 2256	.033 333 333
31	9 61	5.567 7644	.032 258 065
32	10 24	5.656 8542	.031 250 000
33	10 89	5.744 5626	.030 303 030
34	11 56	5.830 9519	.029 411 765
35	12 25	5.916 0798	.028 571 429
36	12 96	6.000 0000	.027 777 778
37	13 69	6.082 7625	.027 027 027
38	14 44	6.164 4140	.026 315 789
39	15 21	6.244 9980	.025 641 026
40	16 00	6.324 5553	.025 000 000
41	16 81	6.403 1242	.024 390 244
42	17 64	6.480 7407	.023 809 524
43	18 49	6.557 4385	.023 255 814
44	19 36	6.633 2496	.022 727 273
45	20 25	6.708 2039	.022 222 222
46	21 16	6.782 3300	.021 739 130
47	22 09	6.855 6546	.021 276 596
48	23 04	6.928 2032	.020 833 333
49	24 01	7.000 0000	.020 408 163
50	25 00	7.071 0678	.020 000 000

¹ Reproduced from *A Manual of Problems and Tables in Statistics* by Mills and Davenport, with special permission of the publishers, Henry Holt and Co.

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
51	26 01	7.141 4284	.019 607 843
52	27 04	7.211 1026	.019 230 769
53	28 09	7.280 1099	.018 867 925
54	29 16	7.348 4692	.018 518 519
55	30 25	7.416 1985	.018 181 818
56	31 36	7.483 3148	.017 857 143
57	32 49	7.549 8344	.017 543 860
58	33 64	7.615 7731	.017 241 379
59	34 81	7.681 1457	.016 949 151
60	36 00	7.745 9667	.016 666 667
61	37 21	7.810 2497	.016 393 443
62	38 44	7.874 0079	.016 129 032
63	39 69	7.937 2539	.015 873 016
64	40 96	8.000 0000	.015 625 000
65	42 25	8.062 2577	.015 384 615
66	43 56	8.124 0384	.015 151 515
67	44 89	8.185 3528	.014 925 373
68	46 24	8.246 2113	.014 705 882
69	47 61	8.306 6239	.014 492 754
70	49 00	8.366 6003	.014 285 714
71	50 41	8.426 1498	.014 084 507
72	51 84	8.485 2814	.013 888 889
73	53 29	8.544 0037	.013 698 630
74	54 76	8.602 3253	.013 513 514
75	56 25	8.660 2540	.013 333 333
76	57 76	8.717 7979	.013 157 895
77	59 29	8.774 9644	.012 987 013
78	60 84	8.831 7609	.012 820 513
79	62 41	8.888 1944	.012 658 228
80	64 00	8.944 2719	.012 500 000
81	65 61	9.000 0000	.012 345 679
82	67 24	9.055 3851	.012 195 122
83	68 89	9.110 4336	.012 048 193
84	70 56	9.165 1514	.011 904 762
85	72 25	9.219 5445	.011 764 706
86	73 96	9.273 6185	.011 627 907
87	75 69	9.327 3791	.011 494 253
88	77 44	9.380 8315	.011 363 636
89	79 21	9.433 9811	.011 235 955
90	81 00	9.486 8330	.011 111 111
91	82 81	9.539 3920	.010 989 011
92	84 64	9.591 6630	.010 869 565
93	86 49	9.643 6508	.010 752 688
94	88 36	9.695 3597	.010 638 298
95	90 25	9.746 7943	.010 526 316
96	92 16	9.797 9590	.010 416 667
97	94 09	9.848 8578	.010 309 278
98	96 04	9.899 4949	.010 204 082
99	98 01	9.949 8744	.010 101 010
100	1 00 00	10.000 0000	.010 000 000

**SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000**

n	n^2	$n^{1/2}$	$1/n$
101	1 02 01	10.049 8756	.009 900 990
102	1 04 04	10.099 5049	.009 803 922
103	1 06 09	10.148 8916	.009 708 738
104	1 08 16	10.198 0390	.009 615 385
105	1 10 25	10.246 9508	.009 523 810
106	1 12 36	10.295 6301	.009 433 962
107	1 14 49	10.344 0804	.009 345 794
108	1 16 64	10.392 3048	.009 259 259
109	1 18 81	10.440 3065	.009 174 312
110	1 21 00	10.488 0885	.009 090 909
111	1 23 21	10.535 6538	.009 009 009
112	1 25 44	10.583 0052	.008 928 571
113	1 27 69	10.630 1458	.008 849 558
114	1 29 96	10.677 0783	.008 771 930
115	1 32 25	10.723 8053	.008 695 652
116	1 34 56	10.770 3296	.008 620 690
117	1 36 89	10.816 6538	.008 547 009
118	1 39 24	10.862 7805	.008 474 576
119	1 41 61	10.908 7121	.008 403 361
120	1 44 00	10.954 4512	.008 333 333
121	1 46 41	11.000 0000	.008 264 463
122	1 48 84	11.045 3610	.008 196 721
123	1 51 29	11.090 5365	.008 130 081
124	1 53 76	11.135 5287	.008 064 516
125	1 56 25	11.180 3399	.008 000 000
126	1 58 76	11.224 9722	.007 936 508
127	1 61 29	11.269 4277	.007 874 016
128	1 63 84	11.313 7085	.007 812 500
129	1 66 41	11.357 8167	.007 751 938
130	1 69 00	11.401 7543	.007 692 308
131	1 71 61	11.445 5231	.007 633 588
132	1 74 24	11.489 1253	.007 575 758
133	1 76 89	11.532 5626	.007 518 797
134	1 79 56	11.575 8369	.007 462 687
135	1 82 25	11.618 9500	.007 407 407
136	1 84 96	11.661 9038	.007 352 941
137	1 87 69	11.704 6999	.007 299 270
138	1 90 44	11.747 3401	.007 246 377
139	1 93 21	11.789 8261	.007 194 245
140	1 96 00	11.832 1596	.007 142 857
141	1 98 81	11.874 3422	.007 092 199
142	2 01 64	11.916 3753	.007 042 254
143	2 04 49	11.958 2607	.006 993 007
144	2 07 36	12.000 0000	.006 944 444
145	2 10 25	12.041 5946	.006 896 552
146	2 13 16	12.083 0460	.006 849 315
147	2 16 09	12.124 3557	.006 802 721
148	2 19 04	12.165 5251	.006 756 757
149	2 22 01	12.206 5556	.006 711 409
150	2 25 00	12.247 4487	.006 666 667

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
151	2 28 01	12.288 2057	.006 622 517
152	2 31 04	12.328 8280	.006 578 947
153	2 34 09	12.369 3169	.006 535 948
154	2 37 16	12.409 6736	.006 493 506
155	2 40 25	12.449 8996	.006 451 613
156	2 43 36	12.489 9960	.006 410 256
157	2 46 49	12.529 9641	.006 369 427
158	2 49 64	12.569 8051	.006 329 114
159	2 52 81	12.609 5202	.006 289 306
160	2 56 00	12.649 1106	.006 250 000
161	2 59 21	12.688 5775	.006 211 180
162	2 62 44	12.727 9221	.006 172 840
163	2 65 69	12.767 1453	.006 134 969
164	2 68 96	12.806 2485	.006 097 561
165	2 72 25	12.845 2326	.006 060 606
166	2 75 56	12.884 0987	.006 024 096
167	2 78 89	12.922 8480	.005 988 024
168	2 82 24	12.961 4814	.005 952 381
169	2 85 61	13.000 0000	.005 917 160
170	2 89 00	13.038 4048	.005 882 353
171	2 92 41	13.076 6968	.005 847 953
172	2 95 84	13.114 8770	.005 813 953
173	2 99 29	13.152 9464	.005 780 347
174	3 02 76	13.190 9060	.005 747 126
175	3 06 25	13.228 7566	.005 714 286
176	3 09 76	13.266 4992	.005 681 818
177	3 13 29	13.304 1347	.005 649 718
178	3 16 84	13.341 6641	.005 617 978
179	3 20 41	13.379 0882	.005 586 592
180	3 24 00	13.416 4079	.005 555 556
181	3 27 61	13.453 6240	.005 524 862
182	3 31 24	13.490 7376	.005 494 505
183	3 34 89	13.527 7493	.005 464 481
184	3 38 56	13.564 6600	.005 434 783
185	3 42 25	13.601 4705	.005 405 405
186	3 45 96	13.638 1817	.005 376 344
187	3 49 69	13.674 7943	.005 347 594
188	3 53 44	13.711 3092	.005 319 149
189	3 57 21	13.747 7271	.005 291 005
190	3 61 00	13.784 0488	.005 263 158
191	3 64 81	13.820 2750	.005 235 602
192	3 68 64	13.856 4065	.005 208 333
193	3 72 49	13.892 4440	.005 181 347
194	3 76 36	13.928 3883	.005 154 639
195	3 80 25	13.964 2400	.005 128 205
196	3 84 16	14.000 0000	.005 102 041
197	3 88 09	14.035 6688	.005 076 142
198	3 92 04	14.071 2473	.005 050 505
199	3 96 01	14.106 7360	.005 025 126
200	4 00 00	14.142 1356	.005 000 000

**SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000**

n	n^2	$n^{1/2}$	$1/n$
201	4 04 01	14.177 4469	.004 975 124
202	4 08 04	14.212 6704	.004 950 495
203	4 12 09	14.247 8068	.004 926 108
204	4 16 16	14.282 8569	.004 901 961
205	4 20 25	14.317 8211	.004 878 049
206	4 24 36	14.352 7001	.004 854 369
207	4 28 49	14.387 4946	.004 830 918
208	4 32 64	14.422 2051	.004 807 692
209	4 36 81	14.456 8323	.004 784 689
210	4 41 00	14.491 3767	.004 761 905
211	4 45 21	14.525 8390	.004 739 336
212	4 49 44	14.560 2198	.004 716 981
213	4 53 69	14.594 5195	.004 694 836
214	4 57 96	14.628 7388	.004 672 897
215	4 62 25	14.662 8783	.004 651 163
216	4 66 56	14.696 9385	.004 629 630
217	4 70 89	14.730 9199	.004 608 295
218	4 75 24	14.764 8231	.004 587 156
219	4 79 61	14.798 6486	.004 566 210
220	4 84 00	14.832 3970	.004 545 455
221	4 88 41	14.866 0687	.004 524 887
222	4 92 84	14.899 6644	.004 504 505
223	4 97 29	14.933 1845	.004 484 305
224	5 01 76	14.966 6295	.004 464 286
225	5 06 25	15.000 0000	.004 444 444
226	5 10 76	15.033 2964	.004 424 779
227	5 15 29	15.066 5192	.004 405 286
228	5 19 84	15.099 6689	.004 385 965
229	5 24 41	15.132 7460	.004 366 812
230	5 29 00	15.165 7509	.004 347 826
231	5 33 61	15.198 6842	.004 329 004
232	5 38 24	15.231 5462	.004 310 345
233	5 42 89	15.264 3375	.004 291 845
234	5 47 56	15.297 0585	.004 273 504
235	5 52 25	15.329 7097	.004 255 319
236	5 56 96	15.362 2915	.004 237 288
237	5 61 69	15.394 8043	.004 219 409
238	5 66 44	15.427 2486	.004 201 681
239	5 71 21	15.459 6248	.004 184 100
240	5 76 00	15.491 9334	.004 166 667
241	5 80 81	15.524 1747	.004 149 378
242	5 85 64	15.556 3492	.004 132 231
243	5 90 49	15.588 4573	.004 115 226
244	5 95 36	15.620 4994	.004 098 361
245	6 00 25	15.652 4758	.004 081 633
246	6 05 16	15.684 3871	.004 065 041
247	6 10 09	15.716 2336	.004 048 583
248	6 15 04	15.748 0157	.004 032 258
249	6 20 01	15.779 7338	.004 016 064
250	6 25 00	15.811 3883	.004 000 000

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
251	6 30 01	15.842 9795	.003 984 064
252	6 35 04	15.874 5079	.003 968 254
253	6 40 09	15.905 9737	.003 952 569
254	6 45 16	15.937 3775	.003 937 008
255	6 50 25	15.968 7194	.003 921 569
256	6 55 36	16.000 0000	.003 906 250
257	6 60 49	16.031 2195	.003 891 051
258	6 65 64	16.062 3784	.003 875 969
259	6 70 81	16.093 4769	.003 861 004
260	6 76 00	16.124 5155	.003 846 154
261	6 81 21	16.155 4944	.003 831 418
262	6 86 44	16.186 4141	.003 816 794
263	6 91 69	16.217 2747	.003 802 281
264	6 96 96	16.248 0768	.003 787 879
265	7 02 25	16.278 8206	.003 773 585
266	7 07 56	16.309 5064	.003 759 398
267	7 12 89	16.340 1346	.003 745 318
268	7 18 24	16.370 7055	.003 731 343
269	7 23 61	16.401 2195	.003 717 472
270	7 29 00	16.431 6767	.003 703 704
271	7 34 41	16.462 0776	.003 690 037
272	7 39 84	16.492 4225	.003 676 471
273	7 45 29	16.522 7116	.003 663 004
274	7 50 76	16.552 9454	.003 649 635
275	7 56 25	16.583 1240	.003 636 364
276	7 61 76	16.613 2477	.003 623 188
277	7 67 29	16.643 3170	.003 610 108
278	7 72 84	16.673 3320	.003 597 122
279	7 78 41	16.703 2931	.003 584 229
280	7 84 00	16.733 2005	.003 571 429
281	7 89 61	16.763 0546	.003 558 719
282	7 95 24	16.792 8556	.003 546 099
283	8 00 89	16.822 6038	.003 533 569
284	8 06 56	16.852 2995	.003 521 127
285	8 12 25	16.881 9430	.003 508 772
286	8 17 96	16.911 5345	.003 496 503
287	8 23 69	16.941 0743	.003 484 321
288	8 29 44	16.970 5627	.003 472 222
289	8 35 21	17.000 0000	.003 460 208
290	8 41 00	17.029 3864	.003 448 276
291	8 46 81	17.058 7221	.003 436 426
292	8 52 64	17.088 0075	.003 424 658
293	8 58 49	17.117 2428	.003 412 969
294	8 64 36	17.146 4282	.003 401 361
295	8 70 25	17.175 5640	.003 389 831
296	8 76 16	17.204 6505	.003 378 378
297	8 82 09	17.233 6879	.003 367 003
298	8 88 04	17.262 6765	.003 355 705
299	8 94 01	17.291 6165	.003 344 482
300	9 00 00	17.320 5081	.003 333 333

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
301	9 06 01	17.349 3516	.003 322 259
302	9 12 04	17.378 1472	.003 311 258
303	9 18 09	17.406 8952	.003 300 330
304	9 24 16	17.435 5958	.003 289 474
305	9 30 25	17.464 2492	.003 278 689
306	9 36 36	17.492 8557	.003 267 974
307	9 42 49	17.521 4155	.003 257 329
308	9 48 64	17.549 9288	.003 246 753
309	9 54 81	17.578 3958	.003 236 246
310 ⁰	9 61 00	17.606 8169	.003 225 806
311	9 67 21	17.635 1921	.003 215 434
312	9 73 44	17.663 5217	.003 205 128
313	9 79 69	17.691 8060	.003 194 888
314	9 85 96	17.720 0451	.003 184 713
315	9 92 25	17.748 2393	.003 174 603
316	9 98 56	17.776 3888	.003 164 557
317	10 04 89	17.804 4938	.003 154 574
318	10 11 24	17.832 5545	.003 144 654
319	10 17 61	17.860 5711	.003 134 796
320	10 24 00	17.888 5438	.003 125 000
321	10 30 41	17.916 4729	.003 115 265
322	10 36 84	17.944 3584	.003 105 590
323	10 43 29	17.972 2008	.003 095 975
324	10 49 76	18.000 0000	.003 086 420
325	10 56 25	18.027 7564	.003 076 923
326	10 62 76	18.055 4701	.003 067 485
327	10 69 29	18.083 1413	.003 058 104
328	10 75 84	18.110 7703	.003 048 780
329	10 82 41	18.138 3571	.003 039 514
330	10 89 00	18.165 9021	.003 030 303
331	10 95 61	18.193 4054	.003 021 148
332	11 02 24	18.220 8672	.003 012 048
333	11 08 89	18.248 2876	.003 003 003
334	11 15 56	18.275 6669	.002 994 012
335	11 22 25	18.303 0052	.002 985 075
336	11 28 96	18.330 3028	.002 976 190
337	11 35 69	18.357 5598	.002 967 359
338	11 42 44	18.384 7763	.002 958 580
339	11 49 21	18.411 9526	.002 949 853
340	11 56 00	18.439 0889	.002 941 176
341	11 62 81	18.466 1853	.002 932 551
342	11 69 64	18.493 2420	.002 923 977
343	11 76 49	18.520 2592	.002 915 452
344	11 83 36	18.547 2370	.002 906 977
345	11 90 25	18.574 1756	.002 898 551
346	11 97 16	18.601 0752	.002 890 173
347	12 04 09	18.627 9360	.002 881 844
348	12 11 04	18.654 7581	.002 873 563
349	12 18 01	18.681 5417	.002 865 330
350	12 25 00	18.708 2869	.002 857 143

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
351	12 32 01	18.734 9940	.002 849 003
352	12 39 04	18.761 6630	.002 840 909
353	12 46 09	18.788 2942	.002 832 861
354	12 53 16	18.814 8877	.002 824 859
355	12 60 25	18.841 4437	.002 816 901
356	12 67 36	18.867 9623	.002 808 989
357	12 74 49	18.894 4436	.002 801 120
358	12 81 64	18.920 8879	.002 793 296
359	12 88 81	18.947 2953	.002 785 515
360	12 96 00	18.973 6660	.002 777 7/8
361	13 03 21	19.000 0000	.002 770 083
362	13 10 44	19.026 2976	.002 762 431
363	13 17 69	19.052 5589	.002 754 821
364	13 24 96	19.078 7840	.002 747 253
365	13 32 25	19.104 9732	.002 739 726
366	13 39 56	19.131 1265	.002 732 240
367	13 46 89	19.157 2441	.002 724 796
368	13 54 24	19.183 3261	.002 717 391
369	13 61 61	19.209 3727	.002 710 027
370	13 69 00	19.235 3841	.002 702 703
371	13 76 41	19.261 3603	.002 695 418
372	13 83 84	19.287 3015	.002 688 172
373	13 91 29	19.313 2079	.002 680 965
374	13 98 76	19.339 0796	.002 673 797
375	14 06 25	19.364 9167	.002 666 667
376	14 13 76	19.390 7194	.002 659 574
377	14 21 29	19.416 4878	.002 652 520
378	14 28 84	19.442 2221	.002 645 503
379	14 36 41	19.467 9223	.002 638 522
380	14 44 00	19.493 5887	.002 631 579
381	14 51 61	19.519 2213	.002 624 672
382	14 59 24	19.544 8203	.002 617 801
383	14 66 89	19.570 3858	.002 610 966
384	14 74 56	19.595 9179	.002 604 167
385	14 82 25	19.621 4169	.002 597 403
386	14 89 96	19.646 8827	.002 590 674
387	14 97 69	19.672 3156	.002 583 979
388	15 05 44	19.697 7156	.002 577 320
389	15 13 21	19.723 0829	.002 570 694
390	15 21 00	19.748 4177	.002 564 103
391	15 28 81	19.773 7199	.002 557 545
392	15 36 64	19.798 9899	.002 551 020
393	15 44 49	19.824 2276	.002 544 529
394	15 52 36	19.849 4332	.002 538 071
395	15 60 25	19.874 6069	.002 531 646
396	15 68 16	19.899 7487	.002 525 253
397	15 76 09	19.924 8588	.002 518 892
398	15 84 04	19.949 9373	.002 512 563
399	15 92 01	19.974 9844	.002 506 266
400	16 00 00	20.000 0000	.002 500 000

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
401	16 08 01	20.024 9844	.002 493 766
402	16 16 04	20.049 9377	.002 487 562
403	16 24 09	20.074 8599	.002 481 390
404	16 32 16	20.099 7512	.002 475 248
405	16 40 25	20.124 6118	.002 469 136
406	16 48 36	20.149 4417	.002 463 054
407	16 56 49	20.174 2410	.002 457 002
408	16 64 64	20.199 0099	.002 450 980
409	16 72 81	20.223 7484	.002 444 988
410	16 81 00	20.248 4567	.002 439 024
411	16 89 21	20.273 1349	.002 433 090
412	16 97 44	20.297 7831	.002 427 184
413	17 05 69	20.322 4014	.002 421 308
414	17 13 96	20.346 9899	.002 415 459
415	17 22 25	20.371 5488	.002 409 639
416	17 30 56	20.396 0781	.002 403 846
417	17 38 89	20.420 5779	.002 398 082
418	17 47 24	20.445 0483	.002 392 344
419	17 55 61	20.469 4895	.002 386 635
420	17 64 00	20.493 9015	.002 380 952
421	17 72 41	20.518 2845	.002 375 297
422	17 80 84	20.542 6386	.002 369 668
423	17 89 29	20.566 9638	.002 364 066
424	17 97 76	20.591 2603	.002 358 491
425	18 06 25	20.615 5281	.002 352 941
426	18 14 76	20.639 7674	.002 347 418
427	18 23 29	20.663 9783	.002 341 920
428	18 31 84	20.688 1609	.002 336 449
429	18 40 41	20.712 3152	.002 331 002
430	18 49 00	20.736 4414	.002 325 581
431	18 57 61	20.760 5395	.002 320 186
432	18 66 24	20.784 6097	.002 314 815
433	18 74 89	20.808 6520	.002 309 469
434	18 83 56	20.832 6667	.002 304 147
435	18 92 25	20.856 6536	.002 298 851
436	19 00 96	20.880 6130	.002 293 578
437	19 09 69	20.904 5450	.002 288 330
438	19 18 44	20.928 4495	.002 283 105
439	19 27 21	20.952 3268	.002 277 904
440	19 36 00	20.976 1770	.002 272 727
441	19 44 81	21.000 0000	.002 267 574
442	19 53 64	21.023 7960	.002 262 443
443	19 62 49	21.047 5652	.002 257 336
444	19 71 36	21.071 3075	.002 252 252
445	19 80 25	21.095 0231	.002 247 191
446	19 89 16	21.118 7121	.002 242 152
447	19 98 09	21.142 3745	.002 237 136
448	20 07 04	21.166 0105	.002 232 143
449	20 16 01	21.189 6201	.002 227 171
450	20 25 00	21.213 2034	.002 222 222

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
451	20 34 01	21.236 7606	.002 217 295
452	20 43 04	21.260 2916	.002 212 389
453	20 52 09	21.283 7967	.002 207 506
454	20 61 16	21.307 2758	.002 202 643
455	20 70 25	21.330 7290	.002 197 802
456	20 79 36	21.354 1565	.002 192 982
457	20 88 49	21.377 5583	.002 188 184
458	20 97 64	21.400 9346	.002 183 406
459	21 06 81	21.424 2853	.002 178 649
460	21 16 00	21.447 6106	.002 173 913
461	21 25 21	21.470 9106	.002 169 197
462	21 34 44	21.494 1853	.002 164 502
463	21 43 69	21.517 4348	.002 159 827
464	21 52 96	21.540 6592	.002 155 172
465	21 62 25	21.563 8587	.002 150 538
466	21 71 56	21.587 0331	.002 145 923
467	21 80 89	21.610 1828	.002 141 328
468	21 90 24	21.633 3077	.002 136 752
469	21 99 61	21.656 4078	.002 132 196
470	22 09 00	21.679 4834	.002 127 660
471	22 18 41	21.702 5344	.002 123 142
472	22 27 84	21.725 5610	.002 118 644
473	22 37 29	21.748 5632	.002 114 165
474	22 46 76	21.771 5411	.002 109 705
475	22 56 25	21.794 4947	.002 105 263
476	22 65 76	21.817 4242	.002 100 840
477	22 75 29	21.840 3297	.002 096 436
478	22 84 84	21.863 2111	.002 092 050
479	22 94 41	21.886 0686	.002 087 683
480	23 04 00	21.908 9023	.002 083 333
481	23 13 61	21.931 7122	.002 079 002
482	23 23 24	21.954 4984	.002 074 689
483	23 32 89	21.977 2610	.002 070 393
484	23 42 56	22.000 0000	.002 066 116
485	23 52 25	22.022 7155	.002 061 856
486	23 61 96	22.045 4077	.002 057 613
487	23 71 69	22.068 0765	.002 053 388
488	23 81 44	22.090 7220	.002 049 180
489	23 91 21	22.113 3444	.002 044 990
490	24 01 00	22.135 9436	.002 040 816
491	24 10 81	22.158 5198	.002 036 660
492	24 20 64	22.181 0730	.002 032 520
493	24 30 49	22.203 6033	.002 028 398
494	24 40 36	22.226 1108	.002 024 291
495	24 50 25	22.248 5955	.002 020 202
496	24 60 16	22.271 0575	.002 016 129
497	24 70 09	22.293 4968	.002 012 072
498	24 80 04	22.315 9136	.002 008 032
499	24 90 01	22.338 3079	.002 004 008
500	25 00 00	22.360 6798	.002 000 000

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
501	25 10 01	22.383 0293	.001 996 008
502	25 20 04	22.405 3565	.001 992 032
503	25 30 09	22.427 6615	.001 988 072
504	25 40 16	22.449 9443	.001 984 127
505	25 50 25	22.472 2051	.001 980 198
506	25 60 36	22.494 4438	.001 976 285
507	25 70 49	22.516 6605	.001 972 387
508	25 80 64	22.538 8553	.001 968 504
509	25 90 81	22.561 0283	.001 964 637
510	26 01 00	22.583 1796	.001 960 784
511	26 11 21	22 605 3091	.001 956 947
512	26 21 44	22.627 4170	.001 953 125
513	26 31 69	22 649 5033	.001 949 318
514	26 41 96	22 671 5681	.001 945 525
515	26 52 25	22 693 6114	.001 941 748
516	26 62 56	22 715 6334	.001 937 984
517	26 72 89	22 737 6340	.001 934 236
518	26 83 24	22 759 6134	.001 930 502
519	26 93 61	22 781 5715	.001 926 782
520	27 04 00	22 803 5085	.001 923 077
521	27 14 41	22 825 4244	.001 919 386
522	27 24 84	22 847 3193	.001 915 709
523	27 35 29	22 869 1933	.001 912 046
524	27 45 76	22 891 0463	.001 908 397
525	27 56 25	22.912 8785	.001 904 762
526	27 66 76	22 934 6899	.001 901 141
527	27 77 29	22 956 4806	.001 897 533
528	27 87 84	22 978 2506	.001 893 939
529	27 98 41	23 000 0000	.001 890 359
530	28 09 00	23 021 7289	.001 886 792
531	28 19 61	23.043 4372	.001 883 239
532	28 30 24	23.065 1252	.001 879 699
533	28 40 89	23.086 7928	.001 876 173
534	28 51 56	23.108 4400	.001 872 659
535	28 62 25	23.130 0670	.001 869 159
536	28 72 96	23.151 6738	.001 865 672
537	28 83 69	23.173 2605	.001 862 197
538	28 94 44	23.194 8270	.001 858 736
539	29 05 21	23.216 3735	.001 855 288
540	29 16 00	23.237 9001	.001 851 852
541	29 26 81	23.259 4067	.001 848 429
542	29 37 64	23.280 8935	.001 845 018
543	29 48 49	23.302 3604	.001 841 621
544	29 59 36	23.323 8076	.001 838 235
545	29 70 25	23.345 2351	.001 834 862
546	29 81 16	23.366 6429	.001 831 502
547	29 92 09	23.388 0311	.001 828 154
548	30 03 04	23.409 3998	.001 824 818
549	30 14 01	23.430 7490	.001 821 494
550	30 25 00	23.452 0788	.001 818 182

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
551	30 36 01	23.473 3892	.001 814 882
552	30 47 04	23.494 6802	.001 811 594
553	30 58 09	23.515 9520	.001 808 318
554	30 69 16	23.537 2046	.001 805 054
555	30 80 25	23.558 4380	.001 801 802
556	30 91 36	23.579 6522	.001 798 561
557	31 02 49	23.600 8474	.001 795 332
558	31 13 64	23.622 0236	.001 792 115
559	31 24 81	23.643 1808	.001 788 209
560	31 36 00	23.664 3191	.001 785 714
561	31 47 21	23.685 4386	.001 782 531
562	31 58 44	23.706 5392	.001 779 359
563	31 69 69	23.727 6210	.001 776 199
564	31 80 96	23.748 6842	.001 773 050
565	31 92 25	23.769 7286	.001 769 912
566	32 03 56	23.790 7545	.001 766 784
567	32 14 89	23.811 7618	.001 763 668
568	32 26 24	23.832 7506	.001 760 563
569	32 37 61	23.853 7209	.001 757 469
570	32 49 00	23.874 6728	.001 754 386
571	32 60 41	23.895 6063	.001 751 313
572	32 71 84	23.916 5215	.001 748 252
573	32 83 29	23.937 4184	.001 745 201
574	32 94 76	23.958 2971	.001 742 160
575	33 06 25	23.979 1576	.001 739 130
576	33 17 76	24.000 0000	.001 736 111
577	33 29 29	24.020 8243	.001 733 102
578	33 40 84	24.041 6306	.001 730 104
579	33 52 41	24.062 4188	.001 727 116
580	33 64 00	24.083 1891	.001 724 138
581	33 75 61	24.103 9416	.001 721 170
582	33 87 24	24.124 6762	.001 718 213
583	33 98 89	24.145 3929	.001 715 266
584	34 10 56	24.166 0919	.001 712 329
585	34 22 25	24.186 7732	.001 709 402
586	34 33 96	24.207 4369	.001 706 485
587	34 45 69	24.228 0829	.001 703 578
588	34 57 44	24.248 7113	.001 700 680
589	34 69 21	24.269 3222	.001 697 793
590	34 81 00	24.289 9156	.001 694 915
591	34 92 81	24.310 4916	.001 692 047
592	35 04 64	24.331 0501	.001 689 189
593	35 16 49	24.351 5913	.001 686 341
594	35 28 36	24.372 1152	.001 683 502
595	35 40 25	24.392 6218	.001 680 672
596	35 52 16	24.413 1112	.001 677 852
597	35 64 09	24.433 5834	.001 675 042
598	35 76 04	24.454 0385	.001 672 241
599	35 88 01	24.474 4765	.001 669 449
600	36 00 00	24.494 8974	.001 666 667

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
601	36 12 01	24.515 3013	.001 663 894
602	36 24 04	24.535 6883	.001 661 130
603	36 36 09	24.556 0583	.001 658 375
604	36 48 16	24.576 4115	.001 655 629
605	36 60 25	24.596 7478	.001 652 893
606	36 72 36	24.617 0673	.001 650 165
607	36 84 49	24.637 3700	.001 647 446
608	36 96 64	24.657 6560	.001 644 737
609	37 08 81	24.677 9254	.001 642 036
610	37 21 00	24.698 1781	.001 639 344
611	37 33 21	24.718 4142	.001 636 661
612	37 45 44	24.738 6338	.001 633 987
613	37 57 69	24.758 8368	.001 631 321
614	37 69 96	24.779 0234	.001 628 664
615	37 82 25	24.799 1935	.001 626 016
616	37 94 56	24.819 3473	.001 623 377
617	38 06 89	24.839 4847	.001 620 746
618	38 19 24	24.859 6058	.001 618 123
619	38 31 61	24.879 7106	.001 615 509
620	38 44 00	24.899 7992	.001 612 903
621	38 56 41	24.919 8716	.001 610 306
622	38 68 84	24.939 9278	.001 607 717
623	38 81 29	24.959 9679	.001 605 136
624	38 93 76	24.979 9920	.001 602 564
625	39 06 25	25.000 0000	.001 600 000
626	39 18 76	25.019 9920	.001 597 444
627	39 31 29	25.039 9681	.001 594 896
628	39 43 84	25.059 9282	.001 592 357
629	39 56 41	25.079 8724	.001 589 825
630	39 69 00	25.099 8008	.001 587 302
631	39 81 61	25.119 7134	.001 584 786
632	39 94 24	25.139 6102	.001 582 278
633	40 06 89	25.159 4913	.001 579 779
634	40 19 56	25.179 3566	.001 577 287
635	40 32 25	25.199 2063	.001 574 803
636	40 44 96	25.219 0404	.001 572 327
637	40 57 69	25.238 8589	.001 569 859
638	40 70 44	25.258 6619	.001 567 398
639	40 83 21	25.278 4493	.001 564 945
640	40 96 00	25.298 2213	.001 562 500
641	41 08 81	25.317 9778	.001 560 062
642	41 21 64	25.337 7189	.001 557 632
643	41 34 49	25.357 4447	.001 555 210
644	41 47 36	25.377 1551	.001 552 795
645	41 60 25	25.396 8502	.001 550 388
646	41 73 16	25.416 5301	.001 547 988
647	41 86 09	25.436 1947	.001 545 595
648	41 99 04	25.455 8441	.001 543 210
649	42 12 01	25.475 4784	.001 540 832
650	42 25 00	25.495 0976	.001 538 462

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
651	42 38 01	25.514 7016	.001 536 098
652	42 51 04	25.534 2907	.001 533 742
653	42 64 09	25.553 8647	.001 531 394
654	42 77 16	25.573 4237	.001 529 052
655	42 90 25	25.592 9678	.001 526 718
656	43 03 36	25.612 4969	.001 524 390
657	43 16 49	25.632 0112	.001 522 070
658	43 29 64	25.651 5107	.001 519 757
659	43 42 81	25.670 9953	.001 517 441
660	43 56 00	25.690 4652	.001 515 152
661	43 69 21	25.709 9203	.001 512 859
662	43 82 44	25.729 3607	.001 510 574
663	43 95 69	25.748 7864	.001 508 296
664	44 08 96	25.768 1975	.001 506 024
665	44 22 25	25.787 5939	.001 503 759
666	44 35 56	25.806 9758	.001 501 502
667	44 48 89	25.826 3431	.001 499 250
668	44 62 24	25.845 6960	.001 497 006
669	44 75 61	25.865 0343	.001 494 768
670	44 89 00	25.884 3582	.001 492 537
671	45 02 41	25.903 6677	.001 490 313
672	45 15 84	25.922 9628	.001 488 095
673	45 29 29	25.942 2435	.001 485 884
674	45 42 76	25.961 5100	.001 483 680
675	45 56 25	25.980 7621	.001 481 481
676	45 69 76	26.000 0000	.001 479 290
677	45 83 29	26.019 2237	.001 477 105
678	45 96 84	26.038 4331	.001 474 926
679	46 10 41	26.057 6284	.001 472 754
680	46 24 00	26.076 8096	.001 470 588
681	46 37 61	26.095 9767	.001 468 429
682	46 51 24	26.115 1297	.001 466 276
683	46 64 89	26.134 2687	.001 464 129
684	46 78 56	26.153 3937	.001 461 988
685	46 92 25	26.172 5047	.001 459 854
686	47 05 96	26.191 6017	.001 457 726
687	47 19 69	26.210 6848	.001 455 604
688	47 33 44	26.229 7541	.001 453 488
689	47 47 21	26.248 8095	.001 451 379
690	47 61 00	26.267 8511	.001 449 275
691	47 74 81	26.286 8789	.001 447 178
692	47 88 64	26.305 8929	.001 445 087
693	48 02 49	26.324 8932	.001 443 001
694	48 16 36	26.343 8797	.001 440 922
695	48 30 25	26.362 8527	.001 438 849
696	48 44 16	26.381 8119	.001 436 782
697	48 58 09	26.400 7576	.001 434 720
698	48 72 04	26.419 6896	.001 432 665
699	48 86 01	26.438 6081	.001 430 615
700	49 00 00	26.457 5131	.001 428 571

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
701	49 14 01	26.476 4046	.001 426 534
702	49 28 04	26.495 2826	.001 424 501
703	49 42 09	26.514 1472	.001 422 475
704	49 56 16	26.532 9983	.001 420 455
705	49 70 25	26.551 8361	.001 418 440
706	49 84 36	26.570 6605	.001 416 431
707	49 98 49	26.589 4716	.001 414 427
708	50 12 64	26.608 2694	.001 412 429
709	50 26 81	26.627 0539	.001 410 437
710	50 41 00	26.645 8252	.001 408 451
711	50 55 21	26.664 5833	.001 406 470
712	50 69 44	26.683 3281	.001 404 494
713	50 83 69	26.702 0598	.001 402 525
714	50 97 96	26.720 7784	.001 400 560
715	51 12 25	26.739 4839	.001 398 601
716	51 26 56	26.758 1763	.001 396 648
717	51 40 89	26.776 8557	.001 394 700
718	51 55 24	26.795 5220	.001 392 758
719	51 69 61	26.814 1754	.001 390 821
720	51 84 00	26.832 8157	.001 388 889
721	51 98 41	26.851 4432	.001 386 963
722	52 12 84	26.870 0577	.001 385 042
723	52 27 29	26.888 6593	.001 383 126
724	52 41 76	26.907 2481	.001 381 215
725	52 56 25	26.925 8240	.001 379 310
726	52 70 76	26.944 3872	.001 377 410
727	52 85 29	26.962 9375	.001 375 516
728	52 99 84	26.981 4751	.001 373 626
729	53 14 41	27.000 0000	.001 371 742
730	53 29 00	27.018 5122	.001 369 863
731	53 43 61	27.037 0117	.001 367 989
732	53 58 24	27.055 4985	.001 366 120
733	53 72 89	27.073 9727	.001 364 256
734	53 87 56	27.092 4344	.001 362 398
735	54 02 25	27.110 8834	.001 360 544
736	54 16 96	27.129 3199	.001 358 696
737	54 31 69	27.147 7439	.001 356 852
738	54 46 44	27.166 1554	.001 355 014
739	54 61 21	27.184 5544	.001 353 180
740	54 76 00	27.202 9410	.001 351 351
741	54 90 81	27.221 3152	.001 349 528
742	55 05 64	27.239 6769	.001 347 709
743	55 20 49	27.258 0263	.001 345 895
744	55 35 36	27.276 3634	.001 344 086
745	55 50 25	27.294 6881	.001 342 282
746	55 65 16	27.313 0006	.001 340 483
747	55 80 09	27.331 3007	.001 338 688
748	55 95 04	27.349 5887	.001 336 898
749	56 10 01	27.367 8644	.001 335 113
750	56 25 00	27.386 1279	.001 333 333

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
751	56 40 01	27.404 3792	.001 331 558
752	56 55 04	27.422 6184	.001 329 787
753	56 70 09	27.440 8455	.001 328 021
754	56 85 16	27.459 0604	.001 326 260
755	57 00 25	27.477 2633	.001 324 503
756	57 15 36	27.495 4542	.001 322 751
757	57 30 49	27.513 6330	.001 321 004
758	57 45 64	27.531 7998	.001 319 261
759	57 60 81	27.549 9546	.001 317 523
760	57 76 00	27.568 0975	.001 315 789
761	57 91 21	27.586 2284	.001 314 060
762	58 06 44	27.604 3475	.001 312 336
763	58 21 69	27.622 4546	.001 310 616
764	58 36 96	27.640 5499	.001 308 901
765	58 52 25	27.658 6334	.001 307 190
766	58 67 56	27.676 7050	.001 305 483
767	58 82 89	27.694 7648	.001 303 781
768	58 98 24	27.712 8129	.001 302 083
769	59 13 61	27.730 8492	.001 300 390
770	59 29 00	27.748 8739	.001 298 701
771	59 44 41	27.766 8868	.001 297 017
772	59 59 84	27.784 8880	.001 295 337
773	59 75 29	27.802 8775	.001 293 661
774	59 90 76	27.820 8555	.001 291 990
775	60 06 25	27.838 8218	.001 290 323
776	60 21 76	27.856 7766	.001 288 660
777	60 37 29	27.874 7197	.001 287 001
778	60 52 84	27.892 6514	.001 285 347
779	60 68 41	27.910 5715	.001 283 697
780	60 84 00	27.928 4801	.001 282 051
781	60 99 61	27.946 3772	.001 280 410
782	61 15 24	27.964 2629	.001 278 772
783	61 30 89	27.982 1372	.001 277 139
784	61 46 56	28.000 0000	.001 275 510
785	61 62 25	28.017 8515	.001 273 885
786	61 77 96	28.035 6915	.001 272 265
787	61 93 69	28.053 5203	.001 270 648
788	62 09 44	28.071 3377	.001 269 036
789	62 25 21	28.089 1438	.001 267 427
790	62 41 00	28.106 9386	.001 265 823
791	62 56 81	28.124 7222	.001 264 223
792	62 72 64	28.142 4946	.001 262 626
793	62 88 49	28.160 2557	.001 261 034
794	63 04 36	28.178 0056	.001 259 446
795	63 20 25	28.195 7444	.001 257 862
796	63 36 16	28.213 4720	.001 256 281
797	63 52 09	28.231 1884	.001 254 705
798	63 68 04	28.248 8938	.001 253 133
799	63 84 01	28.266 5881	.001 251 564
800	64 00 00	28.284 2712	.001 250 000

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
801	64 16 01	28.301 9434	.001 248 439
802	64 32 04	28.319 6045	.001 246 883
803	64 48 09	28.337 2546	.001 245 330
804	64 64 16	28.354 8938	.001 243 781
805	64 80 25	28.372 5219	.001 242 236
806	64 96 36	28.390 1391	.001 240 695
807	65 12 49	28.407 7454	.001 239 157
808	65 28 64	28.425 3408	.001 237 624
809	65 44 81	28.442 9253	.001 236 094
810	65 61 00	28.460 4989	.001 234 568
811	65 77 21	28.478 0617	.001 233 046
812	65 93 44	28.495 6137	.001 231 527
813	66 09 69	28.513 1549	.001 230 012
814	66 25 96	28.530 6852	.001 228 501
815	66 42 25	28.548 2048	.001 226 994
816	66 58 56	28.565 7137	.001 225 490
817	66 74 89	28.583 2119	.001 223 990
818	66 91 24	28.600 6993	.001 222 494
819	67 07 61	28.618 1760	.001 221 001
820	67 24 00	28.635 6421	.001 219 512
821	67 40 41	28.653 0976	.001 218 027
822	67 56 84	28.670 5424	.001 216 545
823	67 73 29	28.687 9766	.001 215 067
824	67 89 76	28.705 4002	.001 213 592
825	68 06 25	28.722 8132	.001 212 121
826	68 22 76	28.740 2157	.001 210 654
827	68 39 29	28.757 6077	.001 209 190
828	68 55 84	28.774 9891	.001 207 729
829	68 72 41	28.792 3601	.001 206 273
830	68 89 00	28.809 7206	.001 204 819
831	69 05 61	28.827 0706	.001 203 369
832	69 22 24	28.844 4102	.001 201 923
833	69 38 89	28.861 7394	.001 200 480
834	69 55 56	28.879 0582	.001 199 041
835	69 72 25	28.896 3666	.001 197 605
836	69 88 96	28.913 6646	.001 196 172
837	70 05 69	28.930 9523	.001 194 743
838	70 22 44	28.948 2297	.001 193 317
839	70 39 21	28.965 4967	.001 191 895
840	70 56 00	28.982 7535	.001 190 476
841	70 72 81	29.000 0000	.001 189 061
842	70 89 64	29.017 2363	.001 187 648
843	71 06 49	29.034 4623	.001 186 240
844	71 23 36	29.051 6781	.001 184 834
845	71 40 25	29.068 8837	.001 183 432
846	71 57 16	29.086 0791	.001 182 033
847	71 74 09	29.103 2644	.001 180 638
848	71 91 04	29.120 4396	.001 179 245
849	72 08 01	29.137 6046	.001 177 856
850	72 25 00	29.154 7595	.001 176 471

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
851	72 42 01	29.171 9043	.001 175 088
852	72 59 04	29.189 0390	.001 173 709
853	72 76 09	29.206 1637	.001 172 333
854	72 93 16	29.223 2784	.001 170 960
855	73 10 25	29.240 3830	.001 169 591
856	73 27 36	29.257 4777	.001 168 224
857	73 44 49	29.274 5623	.001 166 861
858	73 61 64	29.291 6370	.001 165 501
859	73 78 81	29.308 7018	.001 164 144 ⁶
860	73 96 00	29.325 7566	.001 162 791
861	74 13 21	29.342 8015	.001 161 440
862	74 30 44	29.359 8365	.001 160 093
863	74 47 69	29.376 8616	.001 158 749
864	74 64 96	29.393 8769	.001 157 407
865	74 82 25	29.410 8823	.001 156 069
866	74 99 56	29.427 8779	.001 154 734
867	75 16 89	29.444 8637	.001 153 403
868	75 34 24	29.461 8397	.001 152 074
869	75 51 61	29.478 8059	.001 150 748
870	75 69 00	29.495 7624	.001 149 425
871	75 86 41	29.512 7091	.001 148 106
872	76 03 84	29.529 6461	.001 146 789
873	76 21 29	29.546 5734	.001 145 475
874	76 38 76	29.563 4910	.001 144 165
875	76 56 25	29.580 3989	.001 142 857
876	76 73 76	29.597 2972	.001 141 553
877	76 91 29	29.614 1858	.001 140 251
878	77 08 84	29.631 0648	.001 138 952
879	77 26 41	29.647 9342	.001 137 656
880	77 44 00	29.664 7939	.001 136 364
881	77 61 61	29.681 6442	.001 135 074
882	77 79 24	29.698 4848	.001 133 787
883	77 96 89	29.715 3159	.001 132 503
884	78 14 56	29.732 1375	.001 131 222
885	78 32 25	29.748 9496	.001 129 944
886	78 49 96	29.765 7521	.001 128 668
887	78 67 69	29.782 5452	.001 127 396
888	78 85 44	29.799 3289	.001 126 126
889	79 03 21	29.816 1030	.001 124 859
890	79 21 00	29.832 8678	.001 123 596
891	79 38 81	29.849 6231	.001 122 334
892	79 56 64	29.866 3690	.001 121 076
893	79 74 49	29.883 1056	.001 119 821
894	79 92 36	29.899 8328	.001 118 568
895	80 10 25	29.916 5506	.001 117 318
896	80 28 16	29.933 2591	.001 116 071
897	80 46 09	29.949 9583	.001 114 827
898	80 64 04	29.966 6481	.001 113 586
899	80 82 01	29.983 3287	.001 112 347
900	81 00 00	30.000 0000	.001 111 111

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
901	81 18 01	30.016 6620	.001 109 878
902	81 36 04	30.033 3148	.001 108 647
903	81 54 09	30.049 9584	.001 107 420
904	81 72 16	30.066 5928	.001 106 195
905	81 90 25	30.083 2179	.001 104 972
906	82 08 36	30.099 8339	.001 103 753
907	82 26 49	30.116 4407	.001 102 536
908	82 44 64	30.133 0383	.001 101 322
909	82 62 81	30.149 6269	.001 100 110
910	82 81 00	30.166 2063	.001 098 901
911	82 99 21	30.182 7765	.001 097 695
912	83 17 44	30.199 3377	.001 096 491
913	83 35 69	30.215 8899	.001 095 290
914	83 53 96	30.232 4329	.001 094 092
915	83 72 25	30.248 9669	.001 092 896
916	83 90 56	30.265 4919	.001 091 703
917	84 08 89	30.282 0079	.001 090 513
918	84 27 24	30.298 5148	.001 089 325
919	84 45 61	30.315 0128	.001 088 139
920	84 64 00	30.331 5018	.001 086 957
921	84 82 41	30.347 9818	.001 085 776
922	85 00 84	30.364 4529	.001 084 599
923	85 19 29	30.380 9151	.001 083 424
924	85 37 76	30.397 3683	.001 082 251
925	85 56 25	30.413 8127	.001 081 081
926	85 74 76	30.430 2481	.001 079 914
927	85 93 29	30.446 6747	.001 078 749
928	86 11 84	30.463 0924	.001 077 586
929	86 30 41	30.479 5013	.001 076 426
930	86 49 00	30.495 9014	.001 075 269
931	86 67 61	30.512 2926	.001 074 114
932	86 86 24	30.528 6750	.001 072 961
933	87 04 89	30.545 0487	.001 071 811
934	87 23 56	30.561 4136	.001 070 664
935	87 42 25	30.577 7697	.001 069 519
936	87 60 96	30.594 1171	.001 068 376
937	87 79 69	30.610 4557	.001 067 236
938	87 98 44	30.626 7857	.001 066 098
939	88 17 21	30.643 1069	.001 064 963
940	88 36 00	30.659 4194	.001 063 830
941	88 54 81	30.675 7233	.001 062 699
942	88 73 64	30.692 0185	.001 061 571
943	88 92 49	30.708 3051	.001 060 445
944	89 11 36	30.724 5830	.001 059 322
945	89 30 25	30.740 8523	.001 058 201
946	89 49 16	30.757 1130	.001 057 082
947	89 68 09	30.773 3651	.001 055 966
948	89 87 04	30.789 6086	.001 054 852
949	90 06 01	30.805 8436	.001 053 741
950	90 25 00	30.822 0700	.001 052 632

SQUARES, SQUARE ROOTS AND RECIPROCAL OF THE NATURAL
NUMBERS FROM 1 TO 1000

n	n^2	$n^{1/2}$	$1/n$
951	90 44 01	30.838 2879	.001 051 525
952	90 63 04	30.854 4972	.001 050 420
953	90 82 09	30.870 6981	.001 049 318
954	91 01 16	30.886 8904	.001 048 218
955	91 20 25	30.903 0743	.001 047 120
956	91 39 36	30.919 2497	.001 046 025
957	91 58 49	30.935 4166	.001 044 932
958	91 77 64	30.951 5751	.001 043 841
959	91 96 81	30.967 7251	.001 042 753
960	92 16 00	30.983 8668	.001 041 667
961	92 35 21	31.000 0000	.001 040 583
962	92 54 44	31.016 1248	.001 039 501
963	92 73 69	31.032 2413	.001 038 422
964	92 92 96	31.048 3494	.001 037 344
965	93 12 25	31.064 4491	.001 036 269
966	93 31 56	31.080 5405	.001 035 197
967	93 50 89	31.096 6236	.001 034 126
968	93 70 24	31.112 6984	.001 033 058
969	93 89 61	31.128 7648	.001 031 992
970	94 09 00	31.144 8230	.001 030 928
971	94 28 41	31.160 8729	.001 029 866
972	94 47 84	31.176 9145	.001 028 807
973	94 67 29	31.192 9479	.001 027 749
974	94 86 76	31.208 9731	.001 026 694
975	95 06 25	31.224 9900	.001 025 641
976	95 25 76	31.240 9987	.001 024 590
977	95 45 29	31.256 9992	.001 023 541
978	95 64 84	31.272 9915	.001 022 495
979	95 84 41	31.288 9757	.001 021 450
980	96 04 00	31.304 9517	.001 020 408
981	96 23 61	31.320 9195	.001 019 368
982	96 43 24	31.336 8792	.001 018 330
983	96 62 89	31.352 8308	.001 017 294
984	96 82 56	31.368 7743	.001 016 260
985	97 02 25	31.384 7097	.001 015 228
986	97 21 96	31.400 6369	.001 014 199
987	97 41 69	31.416 5561	.001 013 171
988	97 61 44	31.432 4673	.001 012 146
989	97 81 21	31.448 3704	.001 011 122
990	98 01 00	31.464 2654	.001 010 101
991	98 20 81	31.480 1525	.001 009 082
992	98 40 64	31.496 0315	.001 008 065
993	98 60 49	31.511 9025	.001 007 049
994	98 80 36	31.527 7655	.001 006 036
995	99 00 25	31.543 6206	.001 005 025
996	99 20 16	31.559 4677	.001 004 016
997	99 40 09	31.575 3068	.001 003 009
998	99 60 04	31.591 1380	.001 002 004
999	99 80 01	31.606 9613	.001 001 001
1000	1 00 00 00	31.622 7766	.001 000 000

ORDINATES OF THE NORMAL CURVE¹

Ordinates (y) of the normal probability curve, expressed as percentages of the maximum ordinate (y_0) at the mean.

$$\text{Frequency at the maximum ordinate } (y_0) = \frac{N}{2.5066 \sigma}.$$

The table shows the frequencies for ordinates at a distance $\pm \frac{x}{\sigma}$ from the arithmetic average, expressed as percentages of the maximum ordinate, where x is the deviation from the arithmetic average and σ is the standard deviation.

EXAMPLE. Given $y_0 = 200$, $x = 3.0$, $\sigma = 1.5$, $\frac{x}{\sigma} = 2.0$. Then $y = .13534 \times 200 = 27.068$.

$\frac{x}{\sigma}$	0	1	2	3	4	5	6	7	8	9
0.0	1.00000	99995	99980	99955	99920	99875	99820	99755	99681	99596
0.1	99501	99097	99283	99159	99025	98881	98728	98565	98393	98211
0.2	98020	97819	97609	97390	97161	96923	96676	96421	96156	95882
0.3	95600	95309	95009	94701	94384	94059	93725	93384	93034	92677
0.4	92312	91939	91558	91169	90774	90371	89960	89543	89119	88688
0.5	88250	87805	87354	86897	86433	85963	85488	85006	84518	84025
0.6	83527	83023	82514	82000	81481	80957	80429	79896	79358	78816
0.7	78270	77721	77167	76609	76048	75484	74916	74345	73771	73194
0.8	72615	72033	71448	70861	70272	69680	69087	68492	67896	67297
0.9	66698	66097	65495	64892	64288	63683	63078	62472	61866	61260
1.0	60653	60047	59440	58834	58228	57623	57018	56414	55811	55209
1.1	54607	54007	53409	52811	52215	51621	51028	50437	49848	49260
1.2	48675	48092	47511	46933	46357	45783	45212	44644	44078	43516
1.3	42956	42399	41845	41294	40747	40202	39661	39123	38589	38058
1.4	37531	37007	36488	35971	35459	34950	34445	33944	33447	32954
1.5	32465	31980	31499	31023	30550	30082	29618	29158	28702	28251
1.6	27804	27361	26923	26489	26059	25634	25213	24797	24385	23978
1.7	23575	23176	22782	22392	22007	21627	21250	20879	20511	20148
1.8	19790	19436	19086	18741	18400	18064	17732	17404	17081	16762
1.9	16447	16137	15831	15529	15232	14938	14649	14364	14083	13806
2.0	13534	13265	13000	12740	12483	12230	11982	11737	11496	11258
2.1	11025	10795	10569	10347	10129	9914	9702	9495	9290	9090
2.2	08892	08698	08508	08320	08137	07956	07779	07604	07433	07265
2.3	07101	06939	06780	06624	06471	06321	06174	06030	05888	05750
2.4	05613	05480	05349	05221	05096	04972	04852	04734	04618	04505
2.5	04394	04285	04179	04074	03972	03873	03775	03679	03586	03494
2.6	03405	03317	03232	03148	03066	02986	02908	02831	02757	02683
2.7	02612	02542	02474	02408	02343	02279	02217	02157	02098	02040
2.8	01984	01929	01876	01823	01772	01723	01674	01627	01581	01536
2.9	01492	01449	01408	01367	01328	01289	01252	01215	01179	01145
3.0	01111	01078	01045	01015	00985	00955	00927	00897	00872	00845
3.1	00819	00795	00770	00747	00722	00700	00679	00657	00637	00617
3.2	00598	00579	00561	00541	00526	00509	00491	00476	00461	00446
3.3	00432	00419	00404	00391	00369	00366	00353	00341	00331	00318
3.4	00309	00308	00298	00288	00278	00268	00261	00251	00243	00236
3.	01111	00819	00598	00432	00309	00219	00153	00106	00073	00050
4.	00034	00022	00015	00010	00006	00004	00003	00002	00001	00001
5.	00000									

¹ From Rugg's *Statistical Methods Applied to Education*. Reprinted by permission of, and special arrangement with, the publishers, Houghton Mifflin Co.

AREAS OF THE NORMAL CURVE¹

Percentages of the total frequencies (N) of the normal probability distribution between the arithmetic average and a point $\pm \frac{x}{\sigma}$ from the arithmetic average, where x is the deviation from the arithmetic average and σ is the standard deviation.

EXAMPLE. $N = 752$, $x = 3.0$, $\sigma = 1.5$, $\frac{x}{\sigma} = 2.0$. Then $y = .477250 \times 752 = 358.89$.

$\frac{x}{\sigma}$	0	1	2	3	4	5	6	7	8	9
0.0	000000	003989	007978	011967	015953	019939	023922	027903	031881	035856
0.1	039828	043795	047758	051717	055670	059618	063560	067495	071424	075345
0.2	079260	083166	087064	090954	094835	098706	102568	106420	110261	114092
0.3	117911	121720	125516	129300	133072	136831	140576	144309	148027	151732
0.4	155422	159097	162757	166402	170031	173645	177242	180823	184386	187933
0.5	191463	194974	198468	201944	205402	208840	212260	215661	219043	222405
0.6	225747	229069	232371	235653	238914	242154	245373	248571	251748	254903
0.7	258036	261148	264238	267305	270350	273373	276373	279350	282305	285236
0.8	288145	291030	293892	296731	299546	302338	305106	307850	310570	313267
0.9	315940	318589	321214	323815	326391	328944	331472	333977	336457	338913
1.0	341345	343752	346136	348495	350830	353141	355428	357690	359929	362143
1.1	364334	366501	368643	370762	372857	374928	376976	379000	381000	382977
1.2	384930	386861	388768	390651	392512	394350	396165	397958	399727	401473
1.3	403200	404902	406583	408241	409877	411492	413085	414657	416207	417736
1.4	419243	420730	422196	423642	425066	426471	427855	429219	430563	431888
1.5	433193	434478	435745	436992	438220	439429	440620	441792	442947	444083
1.6	445201	446301	447384	448449	449497	450529	451543	452540	453521	454486
1.7	455435	456367	457284	458185	459071	459941	460796	461636	462462	463273
1.8	464070	464852	465621	466375	467116	467843	468557	469258	469946	470621
1.9	471283	471933	472571	473197	473810	474412	475002	475581	476148	476705
2.0	477250	477784	478308	478822	479325	479818	480301	480774	481237	481691
2.1	482136	482571	482997	483414	483823	484222	484611	484997	485371	485738
2.2	486097	486447	486791	487126	487455	487776	488089	488396	488696	488989
2.3	489276	489556	489830	490097	490358	490613	490863	491106	491341	491576
2.4	491803	492024	492240	492451	492656	492857	493053	493244	493431	493613
2.5	493790	493963	494132	494297	494457	494614	494766	494915	495060	495201
2.6	495339	495473	495604	495731	495855	495975	496093	496207	496319	496427
2.7	496533	496636	496736	496833	496928	497020	497110	497197	497282	497365
2.8	497445	497523	497599	497673	497744	497814	497882	497948	498012	498074
2.9	498134	498193	498250	498305	498359	498411	498462	498511	498559	498605
3.0	498650	498694	498736	498777	498817	498856	498893	498930	498965	498999
3.1	499032	499065	499096	499126	499155	499184	499211	499238	499264	499289
3.2	499313	499336	499359	499381	499402	499423	499443	499462	499481	499499
3.3	499517	499534	499550	499566	499581	499596	499610	499624	499638	499651
3.4	499663	499675	499687	499698	499709	499720	499730	499740	499749	499759
3.5	499767	499776	499784	499792	499800	499807	499815	499822	499828	499835
3.6	499841	499847	499853	499858	499864	499869	499874	499879	499883	499888
3.7	499892	499896	499900	499904	499908	499912	499915	499918	499922	499925
3.8	499928	499931	499933	499936	499939	499941	499943	499946	499948	499950
3.9	499952	499954	499956	499958	499959	499961	499963	499964	499966	499967
4.	499968	499979	499987	499992	499995	499997	499998	499999	499999	499999
5.	499999713									

¹ Modeled after Rugg's *Statistical Methods Applied to Education*. Reprinted by permission of, and special arrangement with, the publishers, Houghton Mifflin Co.

INDEX

A

- Abscissa,
 - definition of, 326
- Accuracy,
 - in a report, 830
 - of statistical data, 26-36
 - counting and measurement, 29-30
 - examples of measurement, 30-32
 - (see Significant figures)
- Adams, T. M., 391
- Adequacy,
 - in a report, 830-31
- Addition, review of, 13-14
- Adjustment of trend,
 - in annual data, 583-86
- Agents, in direct collection,
 - credentials for, 112
 - vs. mail questionnaires, 59-65
 - qualifications of, 111
 - training of, 112
 - use of,
 - bias, 64
 - for complex information, 63
 - for confidential information, 64
 - control of replies, 61-62
 - cost, 62-63
 - if limited area, 60
 - if limited time, 60-61
 - personal element, importance of, 60
 - (see also Schedules and questionnaires)
- Aggregative index,
 - weighted, 472-73
 - formula for, 478
 - without weights, 468-69
 - formula for, 477
- Algebraic treatment,
 - averages, subject to, 431-32
- Amplitude, cyclical,
 - adjustment of, in Annalist Index, 667
 - illustrated, 542
 - method of equalizing variability of, in a
 - business index, 658-59
 - method of measuring, 544
 - treatment of, in Babsonchart, 672
- The Annalist Index of Business Activity, 9
 - construction of, 662-69
 - graph of, 668
- Appearance,
 - of a report, 837
- Appendix,
 - to a report, 840
- Approximate measure of seasonal, 596-99
 - limitations of, 599
 - test for pattern, 597
- Areas of the normal curve,
 - fitted to a given distribution, 766
 - table of, Appendix F
 - (see Normal curve)
- Arithmetic average, 388-405, 429-33
 - calculation of,
 - in frequency distribution,
 - direct method, 399
 - short cut method, 400-404
 - in ungrouped data, 388
 - weighted (ungrouped data) 389-90,
 - 392, 394
 - criteria for selection of, 429-33
 - distinctive features of, 405-6
 - formulas for,
 - in frequency distribution,
 - direct, 399
 - short cut, 400-4
 - in ungrouped data, 389
 - weighted (ungrouped data), 390
 - in frequency distribution, 398-405
 - calculation of, 399-404
 - with open end intervals, 403-4
 - with unequal intervals, 403-4
 - in index number construction, 490-91
 - of ungrouped data, 388-89
 - weighted (ungrouped data), 389-96
 - weighted total, 396-98
 - weights,
 - choice of, 395-96
 - effect of, 391-92
 - total value criterion, 393
- Arrangement,
 - of a report, 834-35
- Array,
 - in preparing frequency distribution, 351-54
- Attribute,
 - quantitative,
 - classification in frequency distribution,
 - 320-21, 350
 - graphic presentation of,
 - (see Frequency distributions, graphs of)
 - as a variable characteristic, 150, 350
 - qualitative,
 - graphic presentation of, 301, 315
 - (see Linear or bar graphs, Pictographs, and Circle graphs)

Attribute—*Cont.*qualitative—*Cont.*

order of classifying, 164

as a variable characteristic, 149-50

ratios, 239

Average daily basis,

in time series,

for adjusting calendar irregularity, 551

Average deviation, 441-44, 448-50, 453-54

actual compared with normal distribution, 449-50

coefficient of variation of, 453

from grouped data, 442-44

formulas for,

directly, 442-43

short method, 443-44

use of median in, 443

from ungrouped data, 441-42

Average of relatives index,

weighted, 475-76

formula for, 478

without weights, 469-71

formula for, 477

Average, type used in index number, 490-93

arithmetic, 490-91

geometric, 492-93

median and mode, 491-92

Averages of calculation, 387-413

(see Arithmetic Average, also Geometric Average)

Averages of position, 387, 413-33

(see Median, also Mode)

Averaging ratios, 265-69

comparability, 265-66

weighting, 266-69

B

The Babsonchart,

construction of, 669-74

graph of, 673

Baldwin, H. C., 670

Band chart, 324-26

Bank clearings,

as business indicator, 650

Bank debits,

as business indicator, 650

definition of, 532

illustration of moving trend method, 627-48

as index of local business conditions, 532

Bar graphs, 315-318, 323

divided bars, 315-17

duo-directional, 316-17

essential features of, 317-18

groups of bars, 315-16

single bars, 299-315

of time series, 323

unbroken scale of, 317-18

Base period,

Central Statistical Board, adopted by, 518-19

of commonly used indexes, 512, 517, 523, 530, 533

of an index number, 483-86

choice of, 483-85

length of, 485-86

Base of ratios,

item to use as, 230-31

number of units expressed in, 231-34

Base shifting, of index numbers, 494-95

Beney, M. Ada, 508, 510, 512

Berner, Robert, 224

Bias,

in direct collection of data, 64

types of, in statistical work, 47

Bibliography,

in a report, 840

Bi-modal distribution, 426-28

Binomial distribution, 753-61

fitted to a given distribution, 760-61

general formula for, 754

 $p \neq q$, 771-75

Pascal triangle, 755

proofs,

 $M = np$, 755-56 $\sigma = \sqrt{npq}$, 756

relation to coin tossing, 752-53

terms as probabilities, 754

values as frequencies, 754

for various values of n , 757-59

adjusting areas, 757-60

Board of Governors of Federal Reserve System,

Index of Department Store Sales, 9, 500

Index of Employment, 500

Index of Industrial Production, 487, 500, 515-22

Bowley, Arthur L., 396, 439, 456

Building contracts,

as business indicator, 652

Bureau of Business Research of Harvard Graduate School of Business Administration, 5

Bureau of Business Research, The Ohio State University,

Index of Bank Debits, Canton, Ohio, 532-35

Index of Employment, 479

Bureau of Business and Social Research of University of Buffalo, 153, 552-53

Index of Business Activity in Buffalo, 676-77

Index of Retail Food Prices, 473

Burgess, W. Randolph, 547

Business,

definition and divisions of, 3

statistical background of, 7-10

- Business data,
 - sources of, 186
 - (see Library sources)
- Business failures,
 - as business indicator, 650
- Business indexes,
 - (see Indexes of business conditions)

C

- Calculation, relative ease of,
 - in averages, 431-32
- Calendar variation,
 - adjustment illustrated, 627
 - data recorded weekly, 554
 - definition of, 548-49
 - methods of adjusting, 551
 - in Annalist Index, 663
 - in Babsonchart, 670
 - overlapping data, 554
- Cantril, Hadley, 86
- Captions, in a table,
 - definition of, 157
- Carli, G. R., 462
- Case investigations, 45-46
- "Cell" table, 726
- Census vs. sample,
 - in planning of investigation, 57-59
- Central tendency, measures of, 387
 - (see Averages)
- Chalkley, Lyman, Jr., 82
- Chambers, George G., 35
- Chance arrangement,
 - effect of, on averages, 431-32
- Characteristics of data, 144-49
 - (see Variables)
- Characteristics of logarithms,
 - rules for determining, 847-48
- Check questions,
 - in schedules and questionnaires, 119
- Checking for accuracy,
 - in constructing graphs, 346-47
 - in tables, 169-70
- Chi-square distribution,
 - diagram of, 778
- Chi-square test, 775-81
 - (see Goodness of fit)
- Circle graphs, 311-13
 - dial indexes, 312-13
 - parts of a total, 311-12
- Clarity of style, in a report, 835
- Clark, Evans, 165
- Class intervals,
 - limits of, 359-61
 - discrete or continuous data, 360-61
 - number of, 356
 - open end,
 - arithmetic average, computation of, 403-4
 - median, determined with, 419
 - mode determined with, 425-26
- Class intervals—*Cont.*
 - unequal,
 - arithmetic average, computation of 403-4
 - equalizing area in, 375-77
 - median, determined with, 419
 - mode, determined with, 425-26
 - width of, 357-8
- Class limits,
 - designation of, 360-61
- Class mark, or midpoint, 359
 - (see Midpoint)
- Classification,
 - changes in, 214-15
 - cross-classification, 155-60
 - definition of, 145
 - elements of, 145-47
 - orders of, 144, 148-50
 - overlapping classes, 147
 - simple, or one-way, 155
 - sub-classification, 163
 - in tabulation, 145-50
- Code sheet, 136-38
- Coding, in mechanical tabulation, 131-36
- Coefficient of correlation, 717-28
 - in a "cell" table, 723-28
 - computation of, 725-28
 - preparation of table, 723-24
 - of ungrouped data, 718-23
 - meaning of, 720
 - preliminary formulas, 718-19
 - product-moment formula, 721-23
- Coefficient of variation, 452-53
- Collection of data,
 - (see Direct sources and Library sources)
 - defining the problem, 53
 - planning the procedure, 56-65
 - agents vs. Mail questionnaires, 59-65
 - (see Agents, in direct collection, and Mail questionnaires)
 - census vs. sample, 57-59
 - (see Sampling)
 - library and direct sources, 56-57
 - (see Direct sources and Library sources)
 - stating the program, 65
 - study of problem, preliminary, 54-55
 - subject,
 - complete statement of, 53
- Columbus, Consumers' Cooperative Association, 358
- Common fractions,
 - addition and subtraction, 18-19
 - division, 19
 - multiplication of, 19
- Comparability,
 - test of, between ratios, 255-60
- Comparisons between ratios,
 - classifications of, 256, 264
- Components of time series,
 - in index numbers, 500
 - outline of, 538
 - (see Time series, components of)

- Composite index numbers,
 formulas for, 477-78
 weighted, 472-76
 aggregative, 472-73
 average of relatives, 475-76
 relative of aggregates, 473-75
 without weights, 467-71
 aggregative, 468-69
 average of relatives, 469-71
 relative of aggregates, 469
- Composite indexes of business, 653-61
 cyclical character of, 653
 problems of construction, 654-61
 amplitude, 658-59
 final index, 659-61
 inverted series, 657-58
 lag, 654-57
 weights, 658
- Comprehension, relative ease of,
 in averages, 430-32
- Conclusions,
 in a report, 840
- Conklin, Maxwell R., 515, 518, 519
- Construction,
 of commonly used indexes, methods of,
 513, 519, 524-25, 530, 533-34
 of index numbers, 464-78
 (see Index numbers, basic methods of
 constructing)
- Continuous data,
 class limits of, 360-61
 definition of, 360
 graph of, 368
- Controlled experiments in sampling,
 (see Statistical regularity, principle of)
- Controlled sampling, 81-89
- Corrected rates,
 in vital statistics, 277
- Correlation, 704-46
 cause and effect relation, 705
 coefficient of, 717-28
 in a "cell" table, 723-28
 formulas, preliminary, 718-19
 product-moment formula, 721-23
 correlation ratio, 732
 definition of, 704
 index of correlation, 732
 partial and multiple, 732
 rank difference measure, 733-34
 recapitulation of formulas, 745-46
 regression line, 706-14
 free hand, 708
 in a "cell" table, 729-31
 least squares method, 708-14
 uses of, 714
 scattergram, 706-7
 second line of regression, 731-32
 standard error of estimate, 714-17
 in a "cell" table, 729
 definition of, 715
- Correlation—*Cont.*
 standard error of estimate—*Cont.*
 form of measuring, 715-16
 meaning of, 716-17
 S_e , 732
 of time series, 735-44
 cautions in using, 742
 computation of, 736-41
 limitations of, 735-36
 use in measuring lag, 742-44
 variables, dependent and independent, 705
- Cost,
 of direct collection of data, 62-63
- Cost of living indexes, 505-14
- Counting and measurement, 29-30
- Cowden, Dudley J., 356
- Credentials,
 for agents in direct collection, 112
- Credit index, 292-93
- Criteria for selecting averages, 429-33
- Cross-classification, 155-60
 (see Classification; Tabulation; Tables,
 statistical)
- Cross-hatching, use of,
 in bar graphs, 315-18
 in circle and sector graphs, 312
 general rule for, 340
 in ratio maps, 307-10
- Cross-reference, use of, 217-19
- Croxton, Frederick E., 356
- Crude rates,
 in vital statistics, 277
- Cumulative frequency graph, 371-74
- Curves,
 based on mathematical functions,
 (see Trend)
 of frequency distributions, 379-81
 of time series line graph, 326
- Cycle of logarithmic scale,
 definition of, 334
- Cyclical analysis of time series,
 method of describing cycles, 616, 620
- Cyclical component,
 in index numbers, 500
 (see Time series, components of)
- Cyclical fluctuations, 615-21
 amplitude illustrated, 542
 complete analysis, 617-19
 four methods described symbolically, 615
 length of, 542-43
 period illustrated, 542
- Cyclical index of business, 659-60

D

- Daily rhythm, in time series, 547
- Data, statistical,
 characteristics of, 144-49
 collection of,
 from direct sources, 92-114
 from library sources, 186-208, 210-26

Data, statistical—*Cont.*

- evaluation of, 219-23
 - understanding the background, 219-20
 - visualizing the collection, 220
- preliminary tabulation and editing of, 118-41
- search for, in libraries, examples of, 224-26
- tabulation of, 144-82
- measures of—*Cont.*
- transcribing of, 223
- verification of, 214-19
 - cross-reference, use of, 217-19
 - discrepancies, kinds and causes of, 214-17

Day, Edmund E., 319

Deciles, 441 • 454

Decimal fractions, 19-20

Definitions,

- in schedules and questionnaires, 100-104

Degrees of freedom,

- t* distribution, 806-11
- variance analysis, 815-21
- z* distribution, 811-15

Department of Agriculture, forms used by, 171-74

Department store operations, analysis of, 291-92

Derivative table,

- plan for, in mechanical tabulation, 140
- (see Tables, types of)

Determinate factors,

- in time series analysis, 548-49

Dimensions of a graph,

- definition of length and width, 339

Direct sources,

- collection of data from, 92-114
 - confidential nature of data, 112
 - schedules and questionnaires, preparation of, 94-111
 - (see Schedules and questionnaires)
 - staff, selection of, 111
 - staff, training of, 112
 - steps in, 93-114
 - supervision of 113-14
- (see Agents, in direct collection; Mail questionnaires)
- definition of, 56, 92

Directness of style, in a report, 835

Discrete data,

- class limits of, 360-61
- definition of, 360
- graph of, 369-71

Discrepancies in data, 214-17

Dispersion,

- definition of, 436
- measures of, 436-54
 - calculated, 441-54
 - (see Average deviation and Standard deviation)
- criteria of, 453-54

Dispersion—*Cont.*

measures of—*Cont.*

- position, 437-41
- (see Range and Quartiles)
- relation between, in normal distribution, 448-52
- relative, 452-53
- (see Variation, coefficients of)
- uses of, 456-58

Division,

- by logarithms, 852-53
- review of, 16

Dodge Corporation, F. W., construction contracts data, 9

Dun and Bradstreet, Weekly Food Price Index, 469

Duning, Raymond W., 292

Duo-directional scale, 316-17, 319

E

Eastman Kodak Co.,

- forms used by, 174-82
- production planning by, 682-702

Editing for preliminary tabulation, 118-21

re-editing, 120-21

steps in, 118-20

Effectiveness in graphs,

- in bars, 318
- planning for, 338-47

Electric power production,

- as a business indicator, 652

Elderton, W. Palin, 771

Employment Indexes, 522-26

Equalizing areas,

- of binomial distributions, 757-60

of unequal frequency classes, 375-77

Extended Median,

- definition of, 414-15

Extensive sampling, 78-81, 88

External investigations, 50-51

Extracting roots,

- by logarithms, 853-54

Extreme items,

- effect on averages, 430, 432

F

Factor reversal test, 493-94

Fairchild Publications Index of Prices of Department Store Goods, 9, 485, 496

Falkner, Helen D., 608

Federal government, chart of organization of, between 192 and 193

Financial statements, analysis of, 292-93

Fisher, Irving, 489, 491, 494

Fisher, R. A., 778, 803, 806, 807, 811, 812, 813

Fisher's Ideal Index Number, 494

Footnotes and references,

- on statistical graphs, 346
- in tables, 162

- Form,**
 of a report, 837-40
 of schedules and questionnaires, 104-11
- Forms, preparation of,**
 for tabulation, 144, 156-67
 (see also Tabular Forms)
- Formulas,**
 for arithmetic average, 389, 390, 399, 402
 for average deviation, 442-44
 for binomial distribution, 754
 for chi-square test, 775
 for coefficients of variation, 453
 for correlation,
 coefficient of, 718, 719, 721, 723, 725, 727, 732, 736, 737, 745-46
 rank difference, 733, 746
 regression line, 709, 712, 713, 730, 731, 745
 standard error of estimate, 715, 716, 729, 732, 745
 for geometric average, 405, 407
 for index number construction, 477-78, 494
 for median, 417-18
 for mode, 423, 424
 for moving trend, 642
 for normal curve, 762
 for partition values, (quartiles, etc.), 439
 for skewness, 455-56
 for standard deviation, 445, 447, 448
 for standard error,
 coefficient of correlation, 802, 803
 differences of two samples, 799
 single sample, 791, 792
 for *t* distribution, 806, 810
 for trend,
 advanced curves, 582
 least squares, 573-82
 parabola, 577-79
 straight line, 573-76
 for *z* distribution, 811
 for *z* transformation, 803-4
- Fractions, 18-20**
 common, 18-19
 decimal, 19-20
- Free-hand trend, 560**
- Frequencies,**
 actual, 362-64
 definition of, 320, 350
 percentage, 364
- Frequency curves, types of, 379-81**
 J-shaped, 380
 normal, 379-80
 skewed, 379-80
 symmetrical, 379-80
 U-shaped, 380
- Frequency distributions, 350-383**
 analysis of, first steps in, 351-55
 arraying the data, 351-353
 on bar diagram, 354
 on tally sheet, 353
- Frequency distributions—*Cont.***
 average of, 398-405
 short method of computation, 400-404
 average deviation of, 442-44
 definition of, 350
 double or "cell", 723
 grouping data, 353-61
 in class intervals, 355
 at individual values, 355
 preliminary, 353-55
 principles for, 355-61
 (see Grouping data in frequency distributions)
 percentage frequencies, 364
 preparation of, example of, 361-64
 quartiles in, 439-40
 standard deviation of, 445-40
- Frequency distributions, graphs of, 364-83**
 comparison of two distributions, 377-79
 construction of, 365-67
 cumulative or ogive, 371-74
 more than and less than, 372-73
 percentage frequencies, 373-74
 frequency polygon, 366-68
 histogram, 366
 Lorenz curve, 381-83
 smooth curve, 365
 unequal classes, adjustment for, 374-77
 uses of, 368-71
 for continuous data, 368-69
 for discrete data, 369-71
- Frequency polygon, 366-68**
- Fundamental operations,**
 in use of numbers, 13-18
 order of performing, 17-1.
- G**
- Gallup poll, description of weighted sampling method, 86-88**
- Geometric average, 405-9, 429-33**
 calculation of,
 in frequency distribution, 408
 in ungrouped data, 405-6
 criteria for selection of, 429-33
 definition of, 405
 distinctive features of, 406, 408-9
 formulas for,
 in frequency distribution, 407
 in ungrouped data, 405
 in frequency distribution, 406-9
 in index number construction, 492
 of ratios, 265
 of ungrouped data, 405-6
- Geometric series,**
 definition of, 337
- Goodness of fit, 775-81**
 chi-square test,
 in binomial distribution, 779-80
 degrees of freedom, 776
 interpretation of results, 777, 780
 meaning of, 775-76
 in normal distribution, 777-9

- Government publications,
selected list of, 197-205
(see Library sources)
- Graphic method,
choice of best, 302
purposes served by, 298-301
- Graphs, statistical, 298-347, 364-83
artistic considerations, 338-41
balance in, 339
border of, 338-40
construction of,
steps in, 302-3, 346-47, 365-67
technical details in, 341-47
contrast in, 340
cross-hatching in, 307-10, 312, 315-18,
340
dimensions of, 339
divisions and grid lines, 343-45
effectiveness in, 302, 338-47
footnotes and references on, 346
key and legend of, 339, 342
labels on, 318, 342-45
lines used in, types of, 340-41
printing on, 339, 341
production planning, use in, 692-97
proportions of, 339
scale, parts of, 342-45
size of, 338
symbols, use of, 300, 301, 341
tests of a good graph, 302
types of,
simple, 303-321
(see Maps, statistical; Circle graphs;
Linear or bar graphs; Pictograms)
two-dimensional, 319-320, 323-38
(see Time series, graphs of; Semi-
logarithmic charts; Frequency dis-
tributions, graphs of; Scatter-
grams)
uses of,
purposes served by, 298-301, 368-71
reasons for, 298
to show non-numerical values and re-
lationships, 300-301
to show numerical values and relation-
ships, 300
- Grid lines,
definition of, 343
- Grouping data in frequency distributions,
class intervals,
limits of, 359-61
number of, 356
width of, 357-59
class limits,
in discrete or continuous data, 360-61
- Guides for items in index, 480-83
location and accessibility of data, 481-82
purpose of index, 480-81
tenets of statistical practice, 482-83
- H
- Hale, Roger F., 174
Hall, L. W., 608
Headings,
in tables, 161-62
of work sheet, preliminary tabulation,
127-29
- Heterogeneous data,
bi-modality, cause of, 426-29
definition of, 74
in tabular form, 144
- Histogram, 366
- Holidays, adjustment for, 551
- Homogeneous data,
definition of, 74
mode of, 426-29
in tabular form, 144
- Hudson, Philip G., 488
Huebner, S. S., 194
- I
- Ideal formula for index number, 494
- Illustrations, list of,
in a report, 838
- Inclusive sampling, 83-85, 89
- Indeterminate factors,
in time series analysis, 549
- "Index" and "index number,"
use of terms, 463
- Index of correlation, 732
- Index numbers, 461-500
basic methods of constructing, 464-78
composite index numbers, 467-78
(see Composite index numbers)
formulas for, 477-78
simple index numbers, examples of,
464-67
comparing, by base shifting, 494-95
construction, problems of, 478
base period, 483-86
(see Base period of index)
items to include guides for, 480-83
(see Guides for items in index)
purpose of index, 479-80
type of average used, 490-93
(see Average, type used in index num-
ber)
weights, 486-90
(see Weights of index)
deflating, by means of, 495-97
graphic comparison by means of, 327-30
graphic presentation of,
dial graphs, 312
(see Time series, graphs of)
historical development of, 462-63
interpretation of, 498, 500
kinds of, 463-64
nomograph, use of, 499-500
tests of 493-95
(see Tests of index numbers)

- Index numbers—*Cont.***
 of time series, comparisons between, 247-49, 255
 unweighted, 467-71, 472, 477
 uses of, 461-62, 495-97
 weighted, 472-78
- Indexes of business conditions, 649-78**
 composite indexes, 653-76
 The Annalist Index, 662-69
 Babsonchart, 669-74
 Buffalo Index of Business Activity, 676-77
 local, 674-77
 method of construction, 653-61
 (see Composite indexes of business)
 cyclical index, 659-60
 form of, 659-61
 single series, 649-52
 trend-cycle index, 660-61
 trend not removed, 661
 trend reinstated, 660
 uses of, 677-78
 as an aid to forecasting, 678
 direct interpretation, 677
- Indexes, commonly used, 506-35**
 bank debits, Canton, Ohio, 532-35
 base period, 533
 method of construction, 533-34
 purpose, 532-33
 sources of data, 533
 cost of living, National Industrial Conference Board, 508-14
 base period, 512
 kinds and sources of data, 509-12
 method of construction, 513
 purpose, 508
 weights, 512-13
 summary, 513-14
 employment, U. S. Bureau of Labor Statistics, 523-26
 base period, 523
 kinds and sources of data, 522-23
 method of construction, 524-25
 purpose, 522
 weights, 523-24
 summary, 526
 industrial production, Federal Reserve Board, 515-22
 base period, 517
 kinds and sources of data, 516-17
 method of construction, 519
 purpose, 516-17
 weights, 520-21
 summary, 521-22
 wholesale price weekly, National Fertilizer Association, 526-31
 base period, 530
 kinds and sources of data, 528-29
 method of construction, 530
 purpose, 527-28
 weights, 530
 summary, 530-31
- Industrial production, indexes of, 514-22**
- Instructions to agents,**
 with schedules, 110-111
- Internal investigations, 49-50**
- International Business Machines Corp.,**
 Electric Accounting Division of, 131-32
- Interquartile range, 440**
 (see **Quartiles**)
- Interruptions in series, 217**
- Intervals,**
 in time series graphs,
 methods of marking, 343-45
 (see **Class intervals**)
- Introduction,**
 to a report, 838-39
- Inverted series in a business index, 657-58**
- Item range in median, 417**
- Items,**
 in terms of ratios, 230-31
 use of all, in averages, 430, 432
- J**
- Jones, Bassett, 462**
- K**
- Katz, Daniel, 86**
- Kendall, M. G., 356, 443, 775, 776, 781, 799**
- Kenney, John F., 443**
- Key,**
 of a graph, 339, 342
- Kinds and sources of data,**
 of commonly used indexes, 509-12, 516-17, 522-23, 528-29, 533
- Known conditions in universe, 76**
- L**
- Labels,**
 location of, on bar graphs, 318
 on scales of graph, 342-43
- Labor turnover,**
 crude rates of, 283
 standardized ratio of, 280-83
- Lag in time series, 654-57**
 adjustment of in Babsonchart, 671
 definition of, 654
 measurement of,
 correlation, 742-44
 tests for, 654-57
 by correlation, 656-57
 graphic, 656
 tabular, 655-56
- Least common denominator, 18-19**
- Least squares trend, 573-79**
 (see **Straight line trend and Parabolic trend**)
- Legend,**
 of a graph, 339, 342

- Letter of transmittal,
 - with questionnaire, 106-9
 - Library and/or direct sources, 56-57
 - (see Library sources and Direct sources)
 - Library sources,
 - classification of, 186-208
 - methods,
 - by form of publication, 188-89
 - by frequency of publication, 189-91
 - by publishing agency, 193-95
 - by regularity of publication, 192-93
 - summary of, 196
 - by types of data, 187-88
 - collection of data from, 186, 210-26
 - meaning of, 186
 - definition of, 56, 186
 - list of, selected,
 - government, 197-205
 - non-governmental, 205-8
 - steps in search for, 210-13
 - examples of, 224-26
 - use of, 210-26
 - evaluation of data, 219-23
 - transcribing data, 223
 - verification of data, 214-19
 - Like items, in ratios,
 - comparison between ratios,
 - kind of relationship, 260-63
 - tests of comparability, 255-60
 - definition of, 237-38
 - Line graphs of time series, 326-38
 - arithmetic scale, 326-32
 - breaks in horizontal scale, 331-32
 - breaks in vertical scale, arithmetic, 330-31
 - comparing several series, 326-330
 - logarithmic scale, 332-38
 - scale equation, 327-30
 - Linear or bar graphs, simple, types of, 313-18
 - (see Bar graphs)
 - Lines, types of,
 - in statistical graphs, 340-41
 - Link-relative measure of seasonal, 603-7
 - steps in process, 606-7
 - test for pattern, 604-6
 - Local indexes of business, 531-35, 674-77
 - advantages and limitations, 675
 - purpose, 674
 - Logarithmic scale, 330, 332
 - advantages of, 332, 335
 - cycle, definition of, 334
 - methods of making, 337-38
 - need for, 330
 - principle of, 332-35
 - relative rates of change, 335-37
 - scale equation in, 335
 - Logarithmic trend, 579-82
 - parabola, 580-82
 - Logarithms,
 - combination of operations, 854-55
 - definition of, 845
 - Logarithms—*Cont.*
 - in division, 852-53
 - explanation of, 845-55
 - in extracting roots, 853-54
 - in multiplication, 852
 - in raising to powers, 853
 - rules for using, 852-55
 - systems of,
 - common or Briggsian, 846
 - natural or Napierian, 762
 - tables of, Appendix C, 857-74
 - two parts of, 846-52
 - characteristics, rules for, 847-48
 - mantissas, use of table of, 849-52
 - use of,
 - in geometric average, computation of, 405-8
 - Lorenz, M. O., 381
 - Lorenz curve, 381-83
- M
- McNair, Malcolm P., 6
 - Mail questionnaires,
 - bias in using, 64
 - cost of using, 62-63
 - definition of, 59
 - for large areas, 60
 - personal element lost in using, 60
 - replies uncontrolled, 61-62
 - for simple information, 63
 - small percentage of replies,
 - reasons for, 61-62
 - time element uncertain, 60-61
 - (see Schedules and Questionnaires)
 - Main text,
 - of a report, 839
 - Mangus, A. R., 443
 - Mantissas, of logarithms,
 - table of, Appendix C, 857-74
 - use of table of, 849-52
 - Maps, location, 303-4
 - Maps, statistical, 303-11
 - definition of, 303
 - flow, 310-11
 - large-dot, 304-6
 - purposes of,
 - to show density, 304
 - to show quantity, 304-6
 - to show ratios, 306-10
 - ratio or cross-hatched, 306-10
 - small or point-dot, 304
 - use of circles of varying size,
 - in flow maps, 310
 - in large-dot maps, 306
 - Measures of central tendency, 387-409, 413-33
 - criteria for selection of, 429-33
 - definition of, 387
 - (see Averages of calculation, and Averages of position)

- Mechanical tabulation, 130-141**
 principles of, 131
 steps in, 131-141
 coding, 131-38
 punching cards, 137
 sorting-counting, 137, 139
 tabulation and cross-tabulation, 139-41
 uses of, 130-31, 141
- Median, 413-20, 429-33**
 criteria for selection of, 429-33
 definition of, 413
 distinctive features of, 429-33
 extended,
 definition of, 415
 use of, 414-15
 formulas for,
 in array, 414
 in frequency distribution, 417-18
 graphic calculation of, 419-20
 in index number construction, 491-92
 item range, definition of, 417
 location of,
 in array, 413-14
 in frequency distribution, 416
 in measures of seasonal variation, 611
 value of,
 in array, 413-16
 in frequency distribution, 416-20
- Methods of sampling, 77-89**
 controlled, 81-88
 definition of, 81
 inclusive, 83-85
 selective, 82-83
 weighted, 85-88
 extensive or uncontrolled, 78-81
- Midpoint or class mark,**
 definition of, 359
- Miscellaneous class, 243-45**
- Mitchell, Wesley C., 462, 487, 492, 493**
- Mode, 413, 420-33**
 in bi-modal distribution, 426-28
 criteria for selection of, 429-33
 definition of, 420
 formulas for, 423, 424
 from grouped data only, 421
 in index number construction, 492
 value of,
 equal class intervals, 421-25
 unequal class intervals, 426-27
- Modley, Rudolf, 315**
- Motives appealed to in letter of transmittal, 106-9**
- Moving-average measure of seasonal, 599-603**
 method of centering, 600
 test for pattern, 602-3
- Moving-average trend, 563-70**
 with actual data, 566-69
 base period,
 with different period cycles, 566-67
 with even number of years, 567-69
 length of, 566
 use of adding machine, 569
- Moving-average trend—Cont.**
 advantages and defects of, 570
 definition of, 563
 fitted to controlled data, 563-66
 keeping current free hand, 569
- Moving seasonal index, 621-23**
- Moving trend, 638-46**
 definition of, 639
 formulas for, 642
 notation described, 642-43
 process described, 643-44
 shift from annual to monthly data, 644-46
- Mudgett, Bruce D., 310**
- Multiple correlation, 732**
- Multiple scale, 319, 326-30**
- Multiplication,**
 by logarithms, 852
 short-cut methods, 15-16
- Mutually exclusive classes,**
 in classification, 146-47
- ## N
- National Fertilizer Association,**
 Weekly Wholesale Price Index, 526-31
- National Industrial Conference Board, 4,**
 27, 499
 Cost of Living Index, 7, 506-14
- Nomograph, 499-500**
- Non-government publications,**
 selected list of, 205-8
 (see Library sources)
- Non-homogeneous data,**
 (see Heterogeneous data)
- "Normal",**
 use of with base period of index, 483-84
 use of with trend, 585
- Normal curve, 379-81, 752-81**
 applications of, 769-71
 equation of, 761-62
 method of fitting, 762-68
 to a given distribution, 765-68
 using the equation, 762-63
 using the table of areas, 763-65
 properties of, 762
 relation to binomial distribution, 753-61
 for actual data, 760-61
 in coin tossing, 753-55
 graphic comparison, 757-59
 relation to probability, 752-53
 relation to skewed distributions, 771-75
- Normal distribution,**
 relation to standard deviation, 451-52
- Normal equations, 573, 709**
- Notation system described, 389, 400, 417,**
 477, 573, 574, 642-43, 715, 753, 775,
 776, 789, 806
- Null-hypothesis, 803-4**
- ## O
- Ogive, 371-74**
- Olenin, Alice, 524, 525**
- Open-end intervals, 360**
 (see Class intervals)

Order of arrangement,
 of bars, in graphs, 315
 of strata, in band chart, 324-26
 Orders of arrangement,
 in tabulation, 144, 148-50, 163-64
 (see Classification, orders of, and Tables,
 statistical arrangement of)
 Ordinate, definition of, 326
 Ordinates of normal curve,
 table of, Appendix E, 895
 equation for, 761-63
 "Other" or "miscellaneous" class, 243-45
 Overlapping classes,
 in classification, 147
 Overlapping items,
 in time ratios, 239-40

P

Parabola trend, 577-79
 discussion of fit, 579
 least squares equations, 577
 Partial correlation, 732
 Partition values, 438-41, 454
 Part to total,
 ratios, percentage distributions, 240-45
 (see Percentage distributions)
 ratios, single, 239-40
 Parts of a total,
 graphic presentation of,
 by circle and sectors, 311-12
 by divided bars, 315-17
 in time series, 324-26
 Pascal triangle, 755
 Pearl, Raymond, 277
 Pearson, Karl, 455, 723, 776, 806
 Pearsonian coefficient of correlation, 723
 Percentage change,
 computation of,
 from logarithmic scale, 334
 use of nomograph in, 499-500
 Percentage difference, 247-50
 Percentage distributions,
 comparisons between, 254-55
 errors to avoid in, 242-45
 graphic presentation of, 311, 316, 317,
 324-26
 "other" or "miscellaneous" class 243-45
 total of, method of rounding-off, 168-69
 Percentage relation, 246-47, 249-50
 Percentages,
 in tabular arrangement, 160, 251
 Percentiles, 441, 454
 Percents, calculation of, 21
 Period,
 cyclical, illustrated, 542
 Phase, positive and negative, of cycle, 542
 Pictograms, 299, 313-15, 341
 Pig-iron production,
 as business indicator, 650

Planning production, 684-702
 product "C", 694-700
 final stage, 698-700
 preliminary stage, 696-98
 product "S", 684-94
 estimating annual sales rate, 691-94
 judgment and objective measurement,
 694
 percent supply required, 684-87
 the planning schedule, 686-91
 Poisson distribution, 775
 Population, statistical,
 definition of, 69
 Precision of style in a report, 835-36
 Preliminary tabulation,
 methods of, 122-41
 mechanical tabulation, 130-41
 (see Mechanical tabulation)
 sorting—counting, 122-23
 tally sheet, 123-25
 work sheet, 125-30
 Presentation of results,
 form of a report, 837-40
 importance of, 826
 requirements of a report, 829-37
 scientific attitude, 827-28
 writer-reader relation, 827-29
 Price cycle, 544-45, 550
 adjustment of, 555-57
 adjustment illustrated, 636
 method of adjusting,
 in the Babson chart, 671
 Price indexes, 463, 465
 as business indicators, 651
 Price level, changes in, 555-57
 Primary table,
 construction of, 152-70
 in mechanical tabulation, 140
 Principle of statistical regularity, 70-76
 (see Statistical regularity)
 Principles of sampling, 785-94
 distribution of samples from a single
 universe, 787-88
 statistical regularity, principles of, 785
 universe and sample, 785-87
 use of a single sample, 788-92
 relation to universe, 788-89
 standard error, 789-92
 universe inferred from sample, 792-94
 Printing on a graph, 339, 341
 Probability, 752-53
 Problem, statement of, in a report, 838
 Product-moment formula for correlation,
 721-28
 algebraic proof, 721
 in a "cell" table, 725-28
 historical argument, 722-23
 ungrouped data, 721-23
 Production planning, 682-702
 examples of, 684-700
 product "C", 694-700
 product "S", 684-94

Production planning—*Cont.*

- general discussion of, 683
- summary and conclusions, 700-2
- Projection of trend, 586-89
- Purchasing power,
 - index numbers of, 496
- Purpose,
 - of commonly used indexes, 508, 516-17, 522, 527-28, 532-33
 - of an index number, 479-80, 481

Q

- Qualitative characteristic,
 - (see Attribute)
- Quantitative characteristic,
 - (see Attribute)
- Quantity indexes, 463-64, 465
- Quartiles, 438-41, 448-50, 453-54
 - definition of, 438
 - formulas for, 439
 - in grouped data, 439-41
 - interquartile range, 440
 - quartile deviation, 440, 453
 - actual compared with normal distribution, 449-50
 - coefficient of variation of, 453
 - semi-interquartile range, 440
 - in ungrouped data, 438-39
- Questionnaire,
 - (see Schedules and questionnaires and Mail questionnaires)
- Questions,
 - sequence of, in schedules and questionnaires, 105
- Quintiles, 441

R

- Railroad freight car loadings,
 - as business indicators, 652
- Railroad ratios, analysis of, 288-90
- Raising to powers,
 - by logarithms, 853
- Random sample, 58, 75
- Range, 437-38
 - from grouped or ungrouped data, 438
- Rank difference correlation, 733-34
- Ratios, 229-95
 - applications of, 274-95
 - averaging of, 265-69
 - (see Averaging ratios)
 - comparisons between, 253-69
 - kinds of, 253-65
 - on different bases, 255-65
 - (see Like items and Unlike items)
 - on same base, 254
 - (see Percentage distributions and Index numbers of time series)
 - (see Averaging ratios and Ratios, applications of)

Ratios—*Cont.*

- compound, 283-88
 - denominator ratio stable, 285-87
 - fluctuating numerator and denominator, 287-88
 - stabilized denominator, 284-85
- construction of, 229-50
 - selection of base, 230-34
 - (see Base, of ratios)
- kinds of, 235-50
 - between like items, 237-50
 - part-to-part and total-to-total, 245-50
 - (see Percentage relation and Percentage difference)
 - part to total, 239-45
 - (see Percentage distribution)
 - between unlike items, 245-37
 - (see Time, Space, and Attribute ratios)
 - presentation of, 250-53
 - including original data, 251-253
 - in tabular form, 250-53
 - in text, 250
 - refined, 274-83
 - illustrative examples of, 274-77
 - standardized, 277-83
 - statistical distinguished from arithmetic, 229-30
 - use of in business, examples of, 288-95
 - department store operations, 291-92
 - financial statements, analysis of, 292-93
 - railroad analysis, 288-90
 - retail credit department analysis, 290-91
 - (see Logarithmic scale and Semi-logarithmic charts)
- Ratio-to-trend measure of seasonal, 608-11
 - steps in process, 608-11
 - test for pattern, 611
 - trend fitted to annual data, 608-9
- Reader, of a report, 828
- Reciprocals,
 - table of, appendix D, 875-94
 - use of in machine division, 466
- References,
 - (see Footnotes and references)
- Refined ratios, 274-83
- Regression line, 707-14
 - in a "cell" table, 729-31
 - equations of, 709-14
 - compared, 713-14
 - origin at (0, 0), 709-11
 - origin at (M_x , M_y), 712-13
 - origin at (M_x , 0), 711-12
 - free hand, 708
- Relation between sample and universe, 792-94
- Relative of aggregates index, 469
 - formula for, 477
- Relative of weighted aggregates index, 473-75
 - formula for, 478
- Relative cycles,
 - argument for using, 583
 - method of computing, 583-84

Reliability, 794-804
 (see Tests of significance)
 Remington Rand Inc., Tabulating Machine
 Division of, 131-32
 Repetition,
 avoidance of, in a report, 836-37
 Replies, percentage of,
 in direct collection of data, 61-62
 Representativeness in sampling, 77-89
 Requirements of a report, 829-37
 accuracy, 830
 adequacy, 830-31
 appearance, 837
 arrangement, 834-35
 soundness, 831-34
 style, 835-36
 Retail credit department analysis, 290-91
 Revisions of series, 215-16
 Rhodes, E. C., 392
 Rigid definition, of averages, 429, 432
 Robinson, A. H., 174, 683
 Rounding-off numbers,
 meaning, in statistical work, 27-28
 rules for, 28-29
 (see Significant figures)
 Running, T. R., 577

S

Sample,
 choice of cases for, in direct collection, 94
 random, 58, 75
 reliability of, 71
 representativeness, methods of securing,
 77-89
 (see Methods of sampling)
 size of, 76-77
 Samples, small, 804-15
 Sampling, 69-89
 importance of, 69-70
 methods of,
 (see Methods of sampling)
 principle of statistical regularity, 70-76
 (see Statistical regularity, principle of)
 principles of,
 (see Principles of sampling)
 problems of, 76-89
 representativeness, 77-89
 size, 76-77
 relation to knowledge, 69
 Saunders, Christopher, 416
 Scale,
 arithmetic, use of in time series graphs,
 326-32
 breaks in,
 of bar graphs, 317-18
 of line graphs of time series,
 horizontal, 331-32
 vertical, 330-31
 double or multiple, 319, 327-30, 335
 duo-directional, 316-17, 319
 location of, on bar graphs, 318

Scale—*Cont.*
 logarithmic, 332-38
 (see Logarithmic scale)
 parts of, in a graph, 342-45
 two-dimensional,
 definition of, 320
 need for, 321
 Scale equation,
 purpose and method, 327-30, 335
 Scattergram, 706-7
 Schedules and questionnaires,
 answers in usable form, 100
 auxiliary material in, 106-11
 instructions to agents, 110-11
 letter of transmittal, 106-9
 bias in answers, 99
 content of, 95-96
 definitions used in, 100-4
 terms, 102
 units, 102-104
 distinction between, 94, 118
 editing of, 118-21
 (see Editing)
 form of, 104-11
 informants and respondents, 98-99
 preparation of, 94-111
 questions, sequence of, 105
 wording of, 97
 Schmalz, Carl N., 226, 290, 291
 Scientific attitude, in report writing, 827-28
 Seasonal component,
 removed from index numbers, 500
 (see Seasonal variation)
 Seasonal pattern,
 changing, 623-24
 definition of, 594
 determining the existence of, 594-95
 fixed, 594
 in index numbers, 462
 Seasonal variation, 592-624
 adjustment illustrated, 638-39
 causes of, 545
 methods of measuring, 595-611
 in Annalist Index, 663
 approximate, 596-99
 link relative, 603-7
 moving average, 599-603
 ratio-to-trend, 608-11
 results of four methods compared,
 612-15
 use of median in, 611-12
 moving seasonal index, 621-23
 nature of, 592-93
 amplitude, 592-93
 regularity, 593
 pattern defined, 547
 seasonal pattern, 594-95
 (see Seasonal pattern)
 Security prices, as business indicators, 651
 Selective sampling, 82-83, 88
 Semi-interquartile range, 440
 (see Quartiles)

- Semi-logarithmic charts**, 329, 335
(see *Logarithmic scale*)
- Shewhart, W. A.**, 786
- Short-cut computation**,
division, 16
multiplication, 15-16
- Short methods of computation**,
average deviation, 443-44
average of frequency distribution, 400-404
by actual deviations, 400-402, 404
in steps, 402-3
standard deviation, 446-48
- Siegfried, Andre**, 189
- Significance**,
(see *Tests of significance*)
- Significant figures**, 27-29
in computation, 32-36
addition, 33-34
division, 35-36
multiplication, 34-35
square root, extracting of, 36
subtraction, 34
definition of, 27
in designation of class limits, 361
number retained in a table, 167-69
rounding off,
method of, 28-29
in percentage distribution, 168-69
- Size**,
of a graph, 338
of a sample, 76-77
- Size groups**,
in ratio maps, 309-310
(see *Class intervals*)
- Skewed curves**, 379-81
of binomial expansions, 771-75
- Skewness**, 454-58
formulas for, 455-56
uses of measures of, 456-58
- Small samples**, 804-15
distinguished from large, 804-5
tests of significance, 805-15
t distribution, 806-11
z distribution, 811-15
- Snyder, Carl**, *Index of General Price Level*, 636
- Sorting, counting**,
in preliminary tabulation, 122-23
- Soundness, in a report**, 831-34
- Sources**,
of commonly used index numbers, 465
- Sources of data**,
library and direct distinguished, 56
(see *Direct sources and Library sources*)
- Space**,
graphic presentation of, 301, 303-311
(see *Maps, statistical*)
order of classifying, 163-64
ratios, 239
as a variable characteristic, 148
- Special purpose indexes**, 464-465
- Square roots, calculation of**, 22-26
- Squares and square roots, table of**, Appendix D, 875-94
- Stabilized denominators**,
in compound ratios, 284-85
- Standard deviation**, 444-54
actual compared with normal distribution, 448-50
coefficient of variation of, 453
definition of, 444
from grouped data, 445-48
formulas for,
directly, 445-46
short method, 446-48
importance of, 451-52
as measure of area of normal curve, 451-52
from ungrouped data,
formulas for, 445
- Standard distribution**,
of department store sales checks, 279-80
of employees, 281-83
- Standard error of estimate**,
in a "cell" table, 729, 732
ungrouped data, 714-17
form of measuring, 715-16
meaning of, 716-17
- Standard error of sampling**,
for single sample, 789-93
definition of, 789
formulas for, 791-92
interpretation of, 789-91
- Standardized ratios**, 277
department store example, 277-80
labor turnover example, 280-83
- Statistical approach**, 1
- Statistical data**,
vs. abstract numbers, 1-2
- Statistical graphs**, 298-347
(see *Graphs, statistical*)
- Statistical investigation**, 43-51
canons of, 46-48
definite object, 46-47
skepticism, 48
unbiased attitude, 47-48
character of, 43-46
external, scope of, 50-51
internal, scope of, 49-50
preliminary planning, 53-65
(see *Collection of data*)
presentation of results of, 826-40
(see *Presentation of results*)
steps in, 48
- Statistical method**, 43-46
definition of, 43-44
use of,
in business, 3
in case investigations, 45-46
in mass investigations, 44-45
- Statistical regularity, principle of**, 70-76
in controlled experiments,
coin tossing, 71-72
dice rolling, 72-73
measuring a desk, 74

Statistical regularity, principle of—*Cont.*

- definition, 70
- in sampling, 785
- with uncontrolled data, 75-76

Statistical series,

- definition of, 319
- types of, 320-21

Statistical source books, 188

- (see Library sources)

Statistical table, example of, 1

- (see Tables, statistical)

Statistical universe, 69

- (see Sampling)

Statistician,

- problems encountered by, 5-7
- type of research undertaken by, 6
- work of, 7-10

Statistics in business, 1-10

Stecker, Margaret Loomis, 508

Steel-ingot production,

- as business indicator, 650

Straight-line trend,

- equation of method of writing, 571-72
- least squares fit, 573-75
- even number of years, 575-76
- odd number of years, 574

Strata chart, 324-26

Stub, in a table,

- definition of, 157

"Student", 806

Sturges, H. A., 356

Style,

- of a report, 835-36

Surface chart, 324-26

Symbols,

- algebraic, in formulas, 389, 400, 417, 477, 573, 574, 642-43, 715, 753, 775, 776, 789, 806
- pictorial in statistical graphs, 300, 301, 341

Symmetrical curves, 379-81

T

t distribution, 806-11

- diagram of, 807

testing significance,

- arithmetic average of a single sample, 806-8
- difference of arithmetic averages, 808-11

Table of contents, in a report, 838

Tables, statistical,

- accompanying graphs, 345-46
- accuracy in, 169-70
- arrangement of,
 - on page, 162-65
 - order of items, 163-64
- captions of, 157
- clarity of wording in,
 - footnotes and references, 162
 - headings, 161-62
 - title, 160

Tables, statistical—*Cont.*

- complexity in, 155-60
- construction of, 152-70
- effectiveness of, 164-66
- errors in, 169-70
- example of, 1
- rulings of, 164-65
- significant figures in, 167-69
- spacing in, 164-65
- stub of, 157
- totals in, 166-68
- type face used in, 164-65
- types of, 150-52
 - derived or derivative,
 - construction of, 154-55, 160, 168-69
 - definition of, 150
 - in mechanical tabulation, 140
 - primary,
 - construction of, 152-70
 - definition of, 150
 - in mechanical tabulation, 140
- unity in, 152-55
- validity of, 169-70

Tabular forms,

- business, 174-82
- government, 171-74
- presentation of ratios in, 250-53
- routine recording, use in, 170

Tabulation, 144-182

- definition of, 144
- in machine tabulation, 139-41
- preliminary, 122-41
- (see Preliminary tabulation)
- preparing schedules for, 121
- types of tables, 150-52
- (see Tables, statistical)

Tally sheet,

- in frequency distribution, 353
- in preliminary tabulation, 123-25

Tarnow, Laurence M., 683

Teele, Stanley F., 746

Testing for seasonal pattern,

- approximate method, 597
- link relative method, 604-6
- moving-average method, 602-3
- ratio-to-trend method, 611

Tests of index numbers, 493-95

- factor reversal, 493-94
- "ideal" formula, 494
- time reversal, 493

Tests of significance, 794-815

- of basic measures, 794-98
 - arithmetic average, 794-96
 - per cent occurrence, 796-98
 - standard deviation, 796
- coefficient of correlation, 802-4
- standard error of r , 802
- the z transformation, 803-4
- two samples from same universe, 798-802
 - difference of arithmetic averages, 799
 - 800

- Tests of significance—*Cont.***
 two samples from same universe—*Cont.*
 difference of per cents of occurrence, 801-2
 difference of standard deviations, 800-1
 Thomas, Woodlief, 515, 518, 519
- Time,**
 graphic presentation of, 301
 (see Time series, graphs of)
 order of classifying, 163-64
 ratios, overlapping and non-overlapping, 239-40
 (see Index numbers of time series)
 as a variable characteristic, 148
- Time reversal test,** 493
- Time series,**
 components of, 537-50
 determinate irregular factors, 548-49
 (see Calendar variation)
 indeterminate irregular factors, 549
 rhythms, 542-48
 (see Seasonal variation, Daily and Weekly rhythms, Price Cycle and Cyclical fluctuations)
 trend, 538-42
 graphs of, 323-38, 539-40
 band, strata or surface, 324-26
 bars, 323
 intervals, methods of marking, 343-45
 lines, 326-38
 (see Line graphs of time series)
- Time series analysis,**
 cyclical, 583-86, 615-21
 preliminary, 550-58
 adjusting calendar irregularity, 550-54
 adjusting price cycle, 555-57
 problem of, 550
 seasonal, 592-624
 trend, 560-89
- Time series correlation,** 735-43
 with cycles, 740-44
 lagged, 742-44
 unlagged, 740-42
 with original data, 736-40
- Title page,** of a report, 837
- Title of tables,** 160
- Tolbert, Lewis E.,** 524, 525
- Total value criterion,** 393
 (see Arithmetic average, weights)
- Totals,**
 included in a table, 166-67
 position of, 167
 rounding-off in a table, 168-69
- Trend,** 538-42, 560-89
 location of, 560-61
 depends on judgment, 561
 fitted, 561
 free hand, 560
 measurement of, 561-83
 flexible and inflexible, 561
- Trend—*Cont.***
 measurement of—*Cont.*
 by mathematical curves,
 logarithmic straight line, 579-82
 parabola, 577-79
 straight line, 571-77
 (see Straight-line trend)
 by moving average, 563-70
 (see Moving-average trend)
 by moving trend, 582, 638-46
 (see Moving trend)
 method of measuring,
 in Annalist Index, 663-66
 in Babsonchart, 673
 projection of, 586-89
 (see Moving trend)
 reason for measuring, 582-89
 for projection, 586-88
 for removal, 583-86
 for separate study, 586
- Trend component,**
 removed from index numbers, 500
 of time series, 538-42, 586
- Trend-cycle indexes of business,** 660-61
- Two-dimensional graphs,** 319-20, 323-38, 364-83, 539-40, 706
- Types of statistical work,** 4-7
 statistical practice, 4-5
- Typographical errors,** 216-17

U

- Uncontrolled sampling,** 78-81, 88
- Unequal intervals,** 358
 (see Class intervals)
- Unfilled orders,** U. S. Steel Corporation
 as business indicator, 651
- Ungrouped data,**
 arithmetic average of, 388-89
- Unit,**
 change in, library data, 214
 definition of, in classification, 146
 in maps, large dot, 304, 306
 in maps, small-dot, 304
 named in title of table, 160
- Units,**
 different, graphic comparison of, 326-30
 fixed and variable, definitions of, 103
 ratios and countable objects, in tabulation, 148
 of scales, on a graph, 343
 used in classification, 146
- United States Bureau of Labor Statistics,**
 Cost of Living Index, 7
 Employment Index, 485, 498, 522-26
 Payrolls Index, 485, 498, 522
 Retail Food Price Index, 479
 Wholesale Price Index, 9, 407-8, 461, 484, 485, 496, 557, 636
- Universe, statistical,**
 definition of, 69

Unlike items, ratios of,
 comparisons between ratios, 263-65
 definition of, 235
 Unweighted index numbers, 467-72
 formulas for, 477
 (see Composite index numbers)
 Use of numbers, 13-36
 (see Fundamental operations, Addition,
 Division, etc., Fractions, Square
 roots and Significant figures)

V

Value indexes, 464, 465
 Variable,
 definition of, 319
 Variable characteristics,
 classification according to, 146
 types of, 148-50
 Variables,
 dependent and independent in correla-
 tion, 705
 definition of, 319-20
 in a frequency distribution, 350
 location of, graphically, 326
 Variance analysis, 815-21
 with agricultural data, 816
 computation illustrated, 817-21
 definition of, 815-16
 limitations of, 817
 use of z distribution, 821
 Variation, coefficients of, 452-53
 definition of, 452
 formulas for, 453
 Veenstra, Theodore A., 84
 Vital statistics,
 rates used in, 233
 crude and corrected, 277

W

Wall, Alexander, 292
 Weekly rhythm, in time series, 547
 Weighted arithmetic average,
 in a frequency distribution, 399-404
 of ungrouped data, 389-96

Weighted index numbers, 472-78
 formulas for, 478
 (see Composite index numbers)
 Weighted sampling, 85-89
 Weighted total, 396-98
 Weights,
 in Annalist Index, 667-68
 in arithmetic average, 395-96
 in averaging ratios, 266-69
 in Babsonchart, 672
 in a business index, 659
 amplitude factor, 659
 judgment factor, 659
 in commonly used indexes, 512-13, 520-
 21, 523-24, 530
 effect of in arithmetic average, 391-95
 of an index, 486-90
 bias due to, 489-90
 constant or variable, 488-89
 physical quantities or values, 487-88
 selection of, 486-87

Wholesale price indexes, 526-31
 Willett, Herbert, 527, 530
 Work sheet, in preliminary tabulation,
 125-30
 Working days, number in month, 554
 Writing reports, 826-40
 (see Requirements of a report)

Y

Yule, G. Udny, 356, 443, 775, 776, 781,
 799

Z

z distribution,
 diagrams of, 812-13
 testing significance,
 difference of two standard deviations,
 814-15
 variance analysis, 821
 z transformation, 803-4
 Zelomek, A. W., 9

