UNIVERSAL
LIBRARY
OU_148883
UNIVERSAL
LIBRARY

OUP—552—7-7-66—10,000

OSMANIA UNIVERSITY LIBRARY

Call No. 311/M65S Accession No. 6220

Author F.C. Mills

Title Statistic. 1 Methods

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

1968

4 AUG 1970

54

American Business Series

GENERAL EDITOR

ROSWELL C. McCREA

Professor of Economics in Columbia University

STATISTICAL METHODS

APPLIED TO ECONOMICS AND BUSINESS

BY

FREDERICK CECIL MILLS

PROFESSOR OF ECONOMICS AND STATISTICS

COLUMBIA UNIVERSITY

REVISED



NEW YORK

HENRY HOLT AND COMPANY

COPYRIGHT, 1924,
BY
HENRY HOLT AND COMPANY
COPYRIGHT, 1938,
BY
HENRY HOLT AND COMPANY, INC.

PRINTED IN THE
UNITED STATES OF AMERICA

PREFACE TO THE REVISED EDITION

During the fourteen years that have elapsed since the first edition of this book was published there has been a very considerable extension of the use of statistical methods in business, in public administration, and in all the social sciences. The pressing requirements of new tasks and new problems, together with increasing knowledge of statistical procedures on the part of administrative and research workers, have contributed to this extension. With this development, the older controversies over qualitative versus quantitative methods have largely been shelved. It is clear that different problems call for different procedures; that the men who are grappling with research problems differ, as regards the methods of analysis they find congenial and fruitful; that induction and deduction are complementary phases of the processes that lead to scientific advance. The choice of research procedures does not necessitate the acceptance of one method and the rejection of another; it calls for the finding of a blend of methods that is adapted to a particular set of problems, and that is suited to the temperament and abilities of the human agent that employs them. For workers dealing with social and economic relations, statistical methods constitute an essential element of this blend. Knowledge of systematic procedures for handling quantitative data, and skill in their use, are necessary parts of the equipment of students of the social sciences and of public and private administrators who must utilize the facts of experience in the formulation of policies.

Gains on this front have been paralleled by notable improvements in statistical techniques. The post-war years have witnessed, in this field, the initiation of such another period of intellectual ferment and creative activity as that which, earlier, brought the great contributions of Karl Pearson and his associates. The older instruments of quantitative analysis have been refined and sharpened; methods of designing statistical experiments and formulating and testing hypotheses have been improved; statistical inference has been placed on a sounder foundation. There can be no doubt that these continuing improvements in the logic and in the technique of statistics will contribute in important ways to the advance of the social sciences and to the betterment of public and private administration.

In preparing the present edition of *Statistical Methods* account has been taken of the more important of the recent developments that have a bearing on the economic and business applications of statistics. In doing this I have sought to retain the main features of the first edition. A systematic development of the fundamentals of statistical method is needed by the beginning student. A working compendium of procedures, with necessary aids to calculation and reference tables, is required by the statistician engaged in administration or research. The book is designed to meet these two needs.

The eighteen chapters of the present edition fall into two main divisions. The first twelve chapters deal with the descriptive aspects of statistics. Induction and sampling are purposely omitted in this development of basic descriptive procedures. Problems of statistical inference, with certain more advanced aspects of statistical description, are discussed in the last six chapters, and in appendices A to E. This organization is, I think, well adapted to the needs of instruction. Some teachers may, indeed, prefer to introduce at an earlier point the concepts of samples and parent populations and the treatment of sampling errors. If so, selected pages from the chapter on elementary probabilities and the normal curve (Chapter XIII) and from the introductory chapter on induction (Chapter XIV) may follow Chapter V in the sequence of study.

In the chapters added to this edition I have sought to exemplify economic applications of the newer methods of analysis. These methods offer rich and, as yet, largely unexplored possibilities to research workers in the social sciences. In these sections I have drawn heavily on the path-breaking work of R. A. Fisher. I am indebted to Dr. Fisher and his publishers, Oliver and Boyd of Edinburgh, for permission to include in this book the tabulations that appear in certain of the Appendix Tables. These, with the other tables included, are designed to make the present book a reasonably complete working manual adapted to the needs of both laboratory worker and student.

I must reaffirm my thanks to those who assisted me in various ways in the preparation of the first edition. I am indebted, in addition, to Jacob M. Gould, Agnes B. Omundson, and William H. Mills for valuable aid in the details of the revision.

F. C. M.

May, 1938.

PREFACE TO THE FIRST EDITION

The last decade has witnessed a remarkable stimulation of interest in quantitative methods in business and in the social sciences. The day when intuition was the chief basis of business judgment and unsupported hypothesis the mode in social studies seems to have passed. Following the lead of workers in the older and traditionally more accurate physical sciences, social scientists and serious students of business are employing in greater measure than ever before a method of study based upon the observation and analysis of *facts*. When these observations are quantitative in character appropriate methods are necessary for their organization and interpretation. This book deals with methods of combining and analyzing such observations, with primary emphasis upon materials drawn from the fields of economics and business.

The justification for limiting the treatment to these particular fields is two-fold. Although general statistical methods are practically universal in their application, special problems are encountered in every field of study. This is particularly true in the realm of economics, which presents many distinctive difficulties and many characteristic problems. Methods that are in some degree specialized to meet these particular requirements have been developed, and these methods call for treatment in a work that is restricted in scope. In the second place, methods can be most effectively explained in terms of particular subjects; abstract methodology is barren of interest to the average person. For these reasons the book has been written with reference to the specific needs of quantitative workers in economics and business.

In the explanation of methods no attempt has been made to secure the brevity of exposition which may be desirable in a strictly mathematical work. The purpose throughout has been to write for the learner not for the finished master, and the explanations have been prepared with the needs of the former in mind. I have felt free to omit certain detailed demonstrations of theorems because this book is presented as an introduction to the subject, not as an exhaustive treatise.

The methods of quantitative analysis that are in general use today represent a long accretion, an accumulation of contributions from workers in many fields. It would be vain to attempt to

enumerate all the individuals who have contributed to the development of the science of statistics. Individual references are given in particular cases in the body of the text, but no list of such acknowledgments can serve as a complete record of the debt modern statisticians owe to their predecessors.

For assistance in the preparation of the material contained in this book I am under many obligations. To Mr. H. E. Anderson and Professor H. B. Killough I am indebted for certain of the data employed in Chapters XI, XVI, and XVII. Professor Warren M. Persons of the Harvard Committee on Economic Research has courteously permitted me to make use of certain results of his work on commodity price index numbers. The index of industrial activity presented in Chapter IX and utilized in Chapter XI is a product of the Statistical Division of the American Telephone and Telegraph Company. I have employed it with the kind permission of Mr. Seymour L. Andrew, Chief Statistician. Suggestions from Professor A. H. Mowbray of the University of California have enabled me to remove several obscurities that were present in an earlier mimeographed edition. I am deeply grateful to Professors Henry L. Moore, Theodore H. Brown, and Henry Schultz for their help in critically reviewing portions of the manuscript. For assistance at every stage of the work involved in the writing of this book I am under deep obligation to Professor Donald H. Davenport. His aid in the collection of material, in the preparation of charts, and in the onerous task of seeing the book through the press has been invaluable. To my wife, above all others, I am indebted for a measure of constant and generous help that cannot be adequately acknowledged here.

F. C. M.

November, 1924.

CONTENTS

CHAPTER		PAGE
I.	Statistical Methods and the Problems of Economics and Business	1
II.	Graphic Presentation	8
III.	The Organization of Statistical Data: The Frequency Distribution	50
IV.	Description of the Frequency Distribution: Averages	86
V.	Description of the Frequency Distribution: Measures of Variation and Skewness	137
VI.	Index Numbers of Prices	161
VII.	The Analysis of Time Series: Measurement of Trend	225
VIII.	The Analysis of Time Series: Measurement of Seasonal and Cyclical Fluctuations	284
IX.	Index Numbers of Physical Volume	305
X.	The Measurement of Relationship: Linear Correlation	325
XI.	The Measurement of Relationship between Time Series	380
XII.	The Measurement of Relationship: Non-Linear Correlation	404
XIII.	Elementary Probabilities and the Normal Curve of Error	425
XIV.	Statistical Induction and the Problem of Sampling	452
XV.	The Analysis of Variance	490
XVI.	The Measurement of Relationship: Multiple and Partial Correlation	531
XVII.	The Measurement of Relationship and the Problem of Estimation	566
XVIII.	Statistical Induction and the Problem of Sampling, Concluded	598

PAGE

APPENDIX:

A. THE METHOD OF LEAST SQUARES AS APPLIED TO CERTAIN STATISTICAL PROBLEMS 638

B. DERIVATION OF FORMULAS FOR MEAN AND STANDARD DEVIATION OF THE BINOMIAL DISTRIBUTION 660

C. DERIVATION OF THE STANDARD ERROR OF THE ARITHMETIC MEAN 664

D. ILLUSTRATING THE MEASUREMENT OF TREND BY A MODIFIED EXPONENTIAL CURVE, A GOMPERTZ CURVE AND A LOGISTIC CURVE 667

E. A FURTHER APPLICATION OF VARIANCE ANALYSIS . 681

F. GLOSSARY OF SYMBOLS 691

APPENDIX TABLE:

I. AREAS OF THE NORMAL CURVE OF ERROR IN TERMS OF ABSCISSA 699

II. TABLE OF t 700

III. VALUES OF THE CORRELATION COEFFICIENT FOR DIFFERENT LEVELS OF SIGNIFICANCE 701

IV. SHOWING THE RELATIONS BETWEEN r AND z FOR VALUES OF z FROM 0 TO 5 702

V. TABLE OF χ^2 703

VI. 1 PER CENT POINTS OF THE DISTRIBUTION OF z . 704

VII. 5 PER CENT POINTS OF THE DISTRIBUTION OF z . 705

VIII. SQUARES OF THE NATURAL NUMBERS FROM 100 TO 999. 706

IX. SUMS OF THE FIRST THREE POWERS OF THE NATURAL NUMBERS FROM 1 TO 50 708

X. TABLE OF 5-PLACE LOGARITHMS OF NUMBERS . . 709

LIST OF REFERENCES 727

INDEX 737

LIST OF CHARTS

FIGURE		PAGE
1.	Location of a Point with Reference to Rectangular Coordinates	9
2.	Production of Passenger Automobiles, by Months, During the Year 1937	11
3.	Graph of the Equation $y = x$	14
4.	Graph of the Equation $y = 2 + 3x$	15
5.	Parabola: Graph of the Equation $y = x^2$	18
6.	Equilateral Hyperbola: Graph of the Equation $y = x^{-1}$	19
7.	Exponential Curve: Graph of the Equation $y = 2^x$	20
8.	Sine Curve: Graph of the Equation $y = \text{Sin } x$	22
9.	Graph of the Equation $\log y = 2 \log x$	27
10.	Graph of the Equation $y = x^2$ (Plotted on paper with logarithmic scales)	29
11.	The Compound Interest Law: Growth of $10.00 at Compound Interest at 6 per cent for 100 Years (Plotted on arithmetic scale)	30
12.	The Compound Interest Law: Growth of $10.00 at Compound Interest at 6 per cent for 100 Years (Plotted on semi-logarithmic or ratio scale)	31
13.	Wheat Flour Exports from the United States, 1913–1936	34
14.	Production of Steel Ingots and Castings in the United States, 1896-1936 (Plotted on semi-logarithmic scale)	35
15.	Exports of the United States, 1920-1936, Showing Total Exports and Exports to Selected Areas	36
16.	Exports of the United States, 1920-1936, Showing Total Exports and Exports to Selected Areas, with Scales of Increase, Decrease, and Comparison	38
17.	Production of Rayon Filament Yarn in the United States, 1912-1937, with Lines Defining Uniform Rates of Growth	40
18.	Farms in New England States in 1935	42
19.	Total Value of Products and Elements of Production	

FIGURE PAGE

Costs, Manufacturing Industries of the United States, 1919–1935 43

20. Comparison of Scheduled and Actual Output (Cumulative) Speedwell Automobile Co., 1936 45

21. Comparison of Scheduled and Actual Output, 1936: Gantt Progress Charts (Showing the situation on February 29th) 47

22. Comparison of Scheduled and Actual Output, 1936: Gantt Progress Chart (Showing the situation on November 30th) 47

23. Column Diagram: Distribution of 210 Employees Classified on the Basis of Weekly Earnings (Class-interval = $2.00) 64

24. Column Diagram: Distribution of 210 Employees Classified on the Basis of Weekly Earnings (Class-interval = $1.00) 65

25. Column Diagram: Distribution of 210 Employees Classified on the Basis of Weekly Earnings (Class-interval = $.50) 66

26. Column Diagram: Distribution of 210 Employees Classified on the Basis of Weekly Earnings (Class-interval = $.25) 66

27. Frequency Polygon: Distribution of 210 Employees Classified on the Basis of Weekly Earnings (Class-interval = $2.00) 67

28. Frequency Polygon: Distribution of 210 Employees Classified on the Basis of Weekly Earnings (Class-interval = $1.00) 67

29. Frequency Polygon: Distribution of 210 Employees Classified on the Basis of Weekly Earnings (Class-interval = $.50) 68

30. Column Diagram: Distribution of Personal Income Recipients in the United States, 1918. Including all Recipients of Incomes below $4,000 (Class-interval = $500) 72

31. Column Diagram: Distribution of Personal Income Recipients in the United States, 1918. Including all Recipients of Incomes below $4,000 (Class-interval $200) 73

FIGURE PAGE

32. Column Diagram: Distribution of Personal Income Recipients in the United States, 1918. Including all Recipients of Incomes below $4,000 (Class-interval $100) 73
33. Frequency Curve: Distribution of Personal Income Recipients in the United States, 1918. Including all Recipients of Incomes below $4,000 (Derived from the column diagram with class-interval of $100) . . . 74
34. Cumulative Frequency Curve: Distribution of Telephone Poles Classified according to Length of Life (Cumulated upward) 80
35. Cumulative Frequency Curve: Distribution of Telephone Poles Classified according to Length of Life (Cumulated downward) 81
36. Distribution of Sawmills in the United States Classified according to Labor Cost in 1921. Illustrating the Structural Relation between the Ogive and the Frequency Curve 83
37. Frequency Curve: Distribution of 18,780 Soldiers Classified according to Height 87
38. Frequency Curve: Distribution of Errors of Observation in Astronomical Measurements 88
39. Zone of Dispersion, Artillery Firing, Showing the Theoretical Percentage Distribution of Shots 89
40. Column Diagram: Distribution of 1,000 Shots from a Single Gun 91
41. Frequency Polygon: Distribution of Heads in a Coin Tossing Experiment 92
42. Frequency Polygon: Distribution of 5,540 Cases of Change in the Wholesale Prices of Commodities from One Year to the Next (after Mitchell) 93
43. Frequency Polygon: Distribution of London-New York Exchange Rates (as recorded over a period of 384 months) 94
44. Distribution of Wage-Earners in Open-Hearth Furnaces, Classified according to Average Weekly Earnings in 1935 95
45. The Normal Curve of Error 98

FIGURE PAGE

46. Illustrating the Location of the Median with Ungrouped Data (Personal incomes of seven individuals) . . . 110
47. Distribution of Weekly Earnings of Employees. A Smoothed Frequency Curve, Showing the Relation between Mean, Median and Mode 122
48. Cumulative Distribution of Weekly Earnings of Employees, Illustrating the Graphic Location of Median and Quartiles 123
49. Frequency Polygon: Distribution of Relative Prices of 670 Commodities in 1927 (Average prices in 1926 = 100) 173
50. Frequency Polygon: Distribution of Relative Prices of 774 Commodities in 1933 (Average prices in 1926 = 100) 174
51. Frequency Polygon: Distribution of Relative Prices of 1,437 Commodities in 1918 (Average prices July 1913 to June 1914 = 100) 177
52. Comparison of Five Simple Index Numbers of Farm Crop Prices, 1919–1935 (1919 = 100) 188
53. Comparison of Four Weighted Index Numbers of Farm Crop Prices, 1919–1935 (1919 = 100) 202
54. New York Clearing House Transactions, 1875–1936, with Moving Averages 233
55. Illustrating the Fitting of a Straight Line to Nine Points 247
56. Illustrating the Fitting of a Second Degree Curve to Nine Points 256
57. Number of Concerns in Business in the United States, 1899–1914, with Line of Trend 259
58. Commercial Failures in the United States, 1897–1933, with Line of Trend 262
59. Production of Petroleum in the United States, 1922–1929, with Line Defining Average Rate of Growth . . 267
60. Production of Crude Petroleum in the United States, 1918–1936, with Line of Trend 268
61. Comparison of Actual and Deflated Values of Contracts Awarded in Engineering Construction, 1913–1936 . 280
62. Frequency Distributions: Monthly Freight Car Loadings Expressed as Relatives of Corresponding Trend Values 289

FIGURE PAGE
63. Freight Car Loadings in the United States, 1918–1927, with Line of Trend and Seasonal Pattern 300
64. Cyclical and Accidental Fluctuations in Freight Car Loadings in the United States, 1918–1927 300
65. Changes in the Physical Volume of Manufacturing Production in the United States, 1914–1935. All Commodities, Capital Goods and Consumption Goods . 308
66. The Growth of Industrial Activity in the United States (1899 = 100) 313
67. Industrial Activity as Related to Long-Term Growth: Percentage Deviations 313
68. Physical Volume of Industrial Production in the United States, 1919–1937 (1923–1925 average = 100) . . 317
69. Scatter Diagram Showing the Relation between Taxable Personal Incomes and Passenger Car Registration, by States, in 1934, with Line of Average Relationship . . 328
70. Tabulation of Items in a Correlation Table 342
71. Scatter Diagram of Federal Reserve and Commercial Bank Rates, with Line of Average Relationship and Zones of Estimate 348
72. Showing the Relation between Discount Rates of Commercial Banks and Federal Reserve Bank Discount Rates 361
73. Showing the Relation between Federal Reserve Bank Discount Rates and the Discount Rates of Commercial Banks 362
74. Showing the Relation between Number of Wage-Earners in Factories and Value of Products in Ten Selected Cities in the State of New York 373
75. Showing the Relation between Number of Wage-Earners in Factories and Value of Products in Eleven Selected Cities in the State of New York 374
76. Cotton Production in the United States, Crop Years 1901–1902 to 1935–1936, with Line of Trend . . . 383
77. Prices of Middling Upland Cotton in New York, Crop Years 1901–1902 to 1935–1936, with Line of Trend . 385
78. Comparison of Cyclical Fluctuations in Industrial Stock Prices and in General Business Activity, 1903–1914 . 391

FIGURE PAGE

79. Coefficients of Correlation between Index of Industrial Stock Prices and Index of Business Activity, 1903–1914, Showing the Results Secured with Different Pairings 393
80. Comparison of Cyclical Fluctuations in Stock Prices and Industrial Activity, 1919–1937 394
81. Coefficients of Correlation between Index of Industrial Stock Prices and Index of Business Activity, 1919–1937, Showing the Results Secured with Different Pairings 397
82. Scatter Diagram Showing the Relation between Alfalfa Yield and Irrigation Water Applied, with Two Lines of Regression 406
83. Scatter Diagram Showing the Relation between Wheat Yield and Nitrogen Applied as Fertilizer, with Straight Line of Regression and Line Joining the Means of the Columns 416
84. A Comparison of Actual and Theoretical Frequencies in a Dice-Rolling Experiment 434
85. An Illustration of the Measurement of Areas under the Normal Curve 438
86. Illustrating the Fitting of a Normal Curve to Frequency Distribution of Telephone Subscribers, Classified according to Message Use 447
87. The Relation between the Production and Price of Oats: Illustrating the Use of an Arithmetic Equation of Regression and Arithmetic Zones of Estimate . . . 591
88. The Relation between the Production and Price of Oats: Illustrating the Use of a Logarithmic Equation of Regression and Geometric Zones of Estimate . . . 592
89. The Relation between the Production and Price of Oats: Illustrating the Use of a Logarithmic Equation of Regression and Geometric Zones of Estimate (Plotted on Double Logarithmic Paper) 593
90. The Relation between the Production and Price of Oats: Illustrating the Use of an Equation of Regression Based upon Reciprocals, and of Harmonic Zones of Estimate 595

FIGURE PAGE

91. Distribution of Standard Deviations in Samples of Four Drawn from a Normal Universe 601

A. Total Production of Rayon Filament Yarn in the United States, 1920–1931, with Modified Exponential Trend 670

B. Total Production of Rayon Filament Yarn in the United States, 1920–1937, with Gompertz Trend Line Extrapolated to 1967 675

C. Railroad Mileage Operated in the United States, by Five-Year Intervals, 1850–1935, with Logistic Trend . 679

STATISTICAL METHODS

CHAPTER I

STATISTICAL METHODS AND THE PROBLEMS OF ECONOMICS AND BUSINESS

The distinction between economics and business rests upon viewpoint and approach, rather than subject matter. The economist and the business man have different objectives, but the substance of the science of economics and the materials with which the art of business administration deals are in large part the same. In this treatise we are concerned with methods that may be employed in handling this common subject matter.

Classes of Business Activity

The tasks that confront business men may, without undue straining, be placed in three classes. First, in logical sequence, are the technical tasks that arise in the processes of production, involving problems of chemistry and physics, of engineering, of animal husbandry, of navigation. The basic technical knowledge called for in the solution of these problems furnishes the foundation of our economic life. This is the domain of the hard-won arts of handling the raw materials and controlling the forces of nature.

In the second class come activities that are connected with the internal organization and administration of individual business units. The technical functions of manipulating organic and inorganic matter for the satisfaction of human wants are performed through administrative units, single farms, mines, factories, railroads, department stores. A whole new division of problems is faced by the business man in organizing these units, in coördinating the work of the different departments, in supervising the daily activities

of the individuals making up each organization. While these are perhaps less fundamental than the technical problems of production, they are, for the average business man, more pressing and more difficult. Scientific method has made less progress in solving these latter problems. There is not the organized body of knowledge which is found in the former field, nor are there the same trained experts to whom the tasks may be delegated.

The two types of economic activity named above include tasks that are in a sense self-centered and controllable. The manufacturer of steel has his technical problems of smelting and refining, his particular administrative duties. The farmer or mine-owner faces the same types of problems, in forms peculiar to his own situation. In the performance of tasks in these fields each man is dealing with problems all the elements of which are under more or less perfect control. Difficulties arise, but these are ordinarily difficulties inherent in the given task, not difficulties arising from a sudden change in the constituent elements of the problem, or the sudden interjection of a new factor. In this respect the third category of tasks to be performed by the business man differs materially from the first two. For this class is composed of problems the elements of which are subject only in part to control by the individuals directly concerned.

This third division includes buying and selling, and all the attendant activities that are carried on in terms of prices. As economic life is at present organized these functions are, to the business man, the most important ones he performs. The technical tasks of production and of internal organization and administration are but means to an end. For the business man the goal of economic activity is the disposal of his product at a profit. The tasks preliminary to this final sale are of necessity subordinated to it, and so performed that the final aim may be achieved. The point of emphasis here is that the business man, in buying and selling, faces problems containing elements which he cannot

control. In securing his raw material, in bringing together the other agents needed in production, and in the final disposal of his product, the business man deals with markets — commodity markets, labor markets, money markets — and finds himself acting in relation to a system of prices quite beyond his control in its major movements. The other less fundamental phases of his activity are subject to a high degree of control, but when the business man comes to the final and most important act, the profitable sale of his product, his power of control dwindles. The motivating force in business activity is the hope of pecuniary profits, pecuniary profits depend upon successful buying and selling, successful buying and selling depend upon favorable conditions in an uncontrollable world of prices — here is the argument that states the major problem of business. And these are the facts which make the price system the dominating and all-important factor in modern business life.

The modern entrepreneur lives in an environment of prices. The term "environment" is not an unapt figure; this world of prices in which the business man functions constitutes a coherent, consistent, well-articulated system of interdependent parts, a system which encompasses all the business activities of the entrepreneur. Since the system is beyond the control of the individual he must adapt himself to it, and must base his activities upon as complete an understanding of the system as he may obtain. Without this understanding the major problems of business are incapable of solution.

Quantitative Character of Economic and Business Problems

Problems falling in the first of the classes outlined above have long been recognized as essentially quantitative in character. Their solution calls for the application of the methods of precision which have been developed in the physical sciences. It is no less true that the strictly eco-

nomic and business problems falling in the other classes require the employment of quantitative methods. Qualitative considerations enter, of course, in the solution of such problems, helping to determine the questions to be asked and the methods to be employed. But facts, measured, weighed and compared with other facts, constitute the basis of business judgments and the foundation of economic reasoning. Statistical methods provide means of organizing and appraising these facts.

Of the three classes of problems distinguished in the preceding section two come within the scope of the present discussion. Though the methods of statistics are in part applicable to the solution of technical problems of production, it is not the purpose of the present work to develop this subject. For the solution of problems in the two other fields — those connected with the internal organization and administration of business units and with the processes of buying and selling that bring the business man into contact with the price system — methods of statistical analysis are peculiarly appropriate.

Statistical Methods and Problems of Internal Administration

The typical business man, in the administration of his organization, is called upon to deal with masses of measurements. He is dealing with tons of coal, cubic feet of gas, or kilowatt hours of energy consumed; with tons of pig iron or pairs of shoes produced; with machine hours and man hours; with wages, costs of production and selling prices expressed in dollars and cents. With the increasing size of the business unit the data with which the administrator must deal become increasingly complicated and numerous, and it becomes increasingly difficult to determine their true significance. Under intuitive or rule-of-thumb methods of administration it is impossible effectively to analyze large masses of data and to control business units above the

average in size. It has been abundantly demonstrated that the law of decreasing returns comes into play in business largely because of administrative difficulties.

Whenever one deals with masses of data the problem is one of condensation and analysis — condensation and simplification in order that it may be possible for limited human faculties to handle the data, analysis (and comparison) in order that the elements of the problem may be distinguished and their significance appreciated. Statistical methods have been developed to facilitate the condensation and analysis of masses of quantitative data.

As a typical example of such a problem may be mentioned the allocation of costs, an operation which has been called cost accounting. The proper analysis of all the factors which enter into this problem is only possible through the use of statistical methods. Accounting methods, restricted to the treatment of pecuniary units, are inadequate for the complete analysis of the items of expense. The analysis of sales records, again, calls for the condensation of masses of data, their representation in simple, understandable form, and their interpretation in relation to other business measurements. The analysis of markets and the study of purchasing records and commodities require the use of quantitative methods not restricted in their application to any one class of measurements. At every hand in internal administration statistical methods may be used to supplement accounting methods, to extend the knowledge of the executive, and to make more effective the control of business operations.

Statistical Methods and External Problems

New problems are encountered when the business man goes into the market to buy or sell. Continually before him are the phenomena of business cycles, and if he is to adapt his producing and marketing policies to the swings of the cycle he must undertake the analysis of these phenomena, employing tools appropriate to the task. Again, the price

system, the movements of which are of such fundamental interest to the business man, requires analysis through the use of quantitative methods. So complex and numerous are the data to be dealt with here that simplification is imperative. Apart somewhat from the immediate interests of the business man, but of dominant importance to the economist, are all the problems connected with the economic process of distribution, the allocation of income and wealth among the agents of production. These, as well as that other great economic problem concerned with the question of value or price determination, are quantitative problems, to be solved through the use of quantitative methods of research.

Statistical Procedures in Research

What are these methods, and wherein does research employing such methods differ from other types of research? Scientific inquiry, whatever its particular method may be, proceeds through careful observation, logical inference and accurate verification. Quantitative methods differ from others only in that observation, inference, and verification are based upon *measurement*. Until measurement is possible in a science it is unavoidable that its observations and findings should lack precision, no matter how brilliant the flashes of intuition nor how painstaking the labors of its students may be. The employment of methods of measurement, making possible the analysis of the factors involved in terms of precise units, gives to a science some of the advantages that sharp-edged tools have over blunt and unreliable instruments. Mathematics and its offspring, statistics and accounting, are the powerful instruments which the modern economist has at his disposal, and of which business, through the development of research agencies and methods, is making constantly greater use.

The tools of the statistician are merely certain mathematical methods, developed for particular types of research. These types of research were not economic in the original

development of statistical methods, but social, political, and anthropometric, with one line of development (that relating to the theory of probabilities) extending back through the field of logic to the gaming table. Yet these tools, developed for work in restricted spheres, have been found to possess much wider applicability, and economics has been one of the newer fields in which the application of these methods, with appropriate alterations and additions, has had fruitful results. The economist has found his hand strengthened and the precision of his work materially increased by the new tools. And business, together with the more abstruse science of economics, has profited.

Reference has been made to the possibility of condensation and simplification through the use of statistical procedures. Such simplification is of cardinal importance in economics and in the other social sciences today. These sciences, to be realistic, must be scrupulously faithful to fact, yet the masses of facts relating to current social processes are, in their magnitude, almost a menace to effective analysis. "Already," writes a reviewer in the *Journal of the Royal Statistical Society*, "economic analysis taxes language to its utmost, and it is a question how much longer mere verbal exposition will be able to control the swelling floods of observable data." Though one may feel that these floods of data fail to provide many of the essential facts about social processes today, there is point to the reviewer's complaint. In the light of this danger systematic procedures in the organization and analysis of data have an importance today that they did not have at an earlier time. Statistical methods constitute such procedures. By their use we may seek to channel and appraise the floods of data, relating to business operations and other social processes, that the fact-gathering agencies of business and government currently release upon us.

CHAPTER II

GRAPHIC PRESENTATION

The explanation of methods of condensing, analyzing, and interpreting the facts of business and economics must start with the discussion of some fundamental considerations which are mathematical rather than statistical in character. In doing so it is deemed advisable, even at the risk of treading quite familiar ground, to explain certain mathematical conceptions to which constant reference will be made in later chapters.

Statistical analysis is concerned primarily with data based upon measurement, expressed either in pecuniary or physical units. The methods of coördinate geometry, developed first by the philosopher Descartes, greatly facilitate the manipulation and interpretation of such data. A summary of the basic principles of coördinate geometry will not be out of place.

Rectangular Coördinates

If two straight lines intersecting each other at right angles are drawn in a plane, it is possible to describe the location of any point in that plane with reference to the point of intersection of the two lines. We will call one of the lines (a vertical line) $Y'Y$, the other line (horizontal) $X'X$, and the point of intersection (or *origin*) O (cf. Fig. 1). If P be any point in the plane, we may draw the line PM, parallel to $Y'Y$ and intersecting $X'X$ at M, and the line PN, parallel to $X'X$ and intersecting $Y'Y$ at N. If we set OM equal to g units and ON equal to h units, g and h constitute the *coördinates* of P, describing its location with reference to the origin O. Thus, in Fig. 1, g equals 6 and h equals 5.

The distance g along the x-axis is termed the *abscissa* of the point P, while the distance h along the y-axis is termed the *ordinate* of the point P. (It is a rule of notation always to give the abscissa first, followed by the ordinate.) The coördinates of any other point in the same plane may be

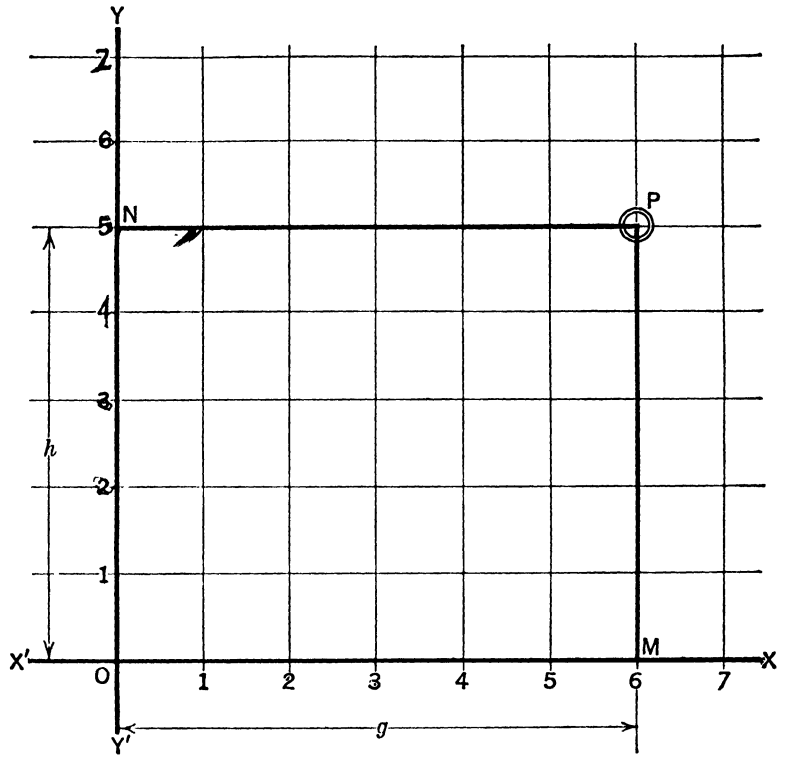


FIG. 1. — Location of a Point with Reference to Rectangular Coördinates

determined in the same way. Conversely, any two real numbers determine a point in the plane, if one be taken as the abscissa and the other as the ordinate.

A point may lie either to the right or left or above or below the origin, O. It is conventional to designate as positive abscissas laid off to the right of the origin, and as negative abscissas laid off to the left of the origin, while

ordinates are positive when laid off above the origin and negative when laid off below the origin. In general, the values to be dealt with in economic statistics lie in the upper right-hand quadrant, where both abscissa and ordinate are positive.

This conception of coördinates is fundamental in mathematics and of basic importance in statistical work. A very simple example will illustrate the utility of this device in representing business data. The figures presented in the following table may be employed.

TABLE 1

Production of Passenger Automobiles in the United States, by Months, During the Year 1937

Month	*Number of passenger cars manufactured*
January	309,637
February	296,636
March	403,879
April	439,980
May	425,432
June	411,394
July	360,403
August	311,456
September	118,671
October	298,662
November	295,328
December	244,385

These data may be represented graphically on the coördinate system, months being laid off along the x-axis and number of automobiles along the y-axis, as in the accompanying diagram (Fig. 2). In plotting the abscissas, December, 1936, is considered as located at the point of origin. The x-value of the entry for January, 1937, is thus 1, of the February figure 2, etc. The coördinates of the point representing the number of cars produced in January, 1937, are 1 and 309,637; for February the values are 2 and 296,636. The coördinates for December are 12 and 244,385. The

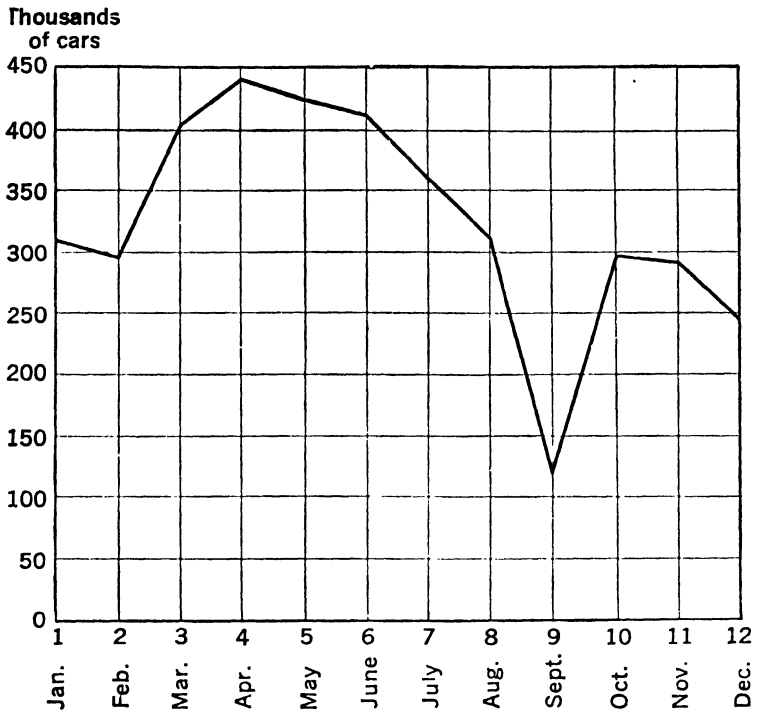


FIG. 2. — Production of Passenger Automobiles, by Months, During the Year 1937

movement of automobile production during the year may be more easily followed if the points are connected by a series of straight lines, as is done in the figure.

INDEPENDENT AND DEPENDENT VARIABLES

In the location of any point by means of coördinates it has been pointed out that two values are involved; every point ties together and expresses a relation between two factors. In the above case these are months and number of passenger automobiles produced. With the passage of time the volume of automobile production changes, and the broken line shows the direction and magnitude of these changes. Both time and number of cars produced are *variables*, that is, they are quantities not of constant value but

characterized by variations in value in the given discussion. Thus in Fig. 1 the abscissa has a fixed value of 6, while the ordinate has a fixed value of 5, but in Fig. 2 both abscissa and ordinate have varying values, the one varying from 1 to 12, the other from 118,671 to 439,980. The symbols x and y are, by convention, used to designate such variable quantities as these, the former in all cases representing the variable plotted along the horizontal axis, the latter representing the variable plotted along the vertical axis.[1]

In Fig. 2, which depicts the changes taking place in automobile production with the passage of time, it will be noted that the latter variable changes by an arbitrary unit, one month. Having made an independent change in the time factor we then determine the change in output taking place during the period thus arbitrarily chopped out. The variable which increases or decreases by increments arbitrarily determined is called the *independent variable*, and is generally plotted on the x-axis. The other variable is termed the *dependent variable*, and is plotted on the y-axis. This dependence may be real, in the sense that the values of the second variable are definitely determined by the values of the independent variable, or it may be purely a conventional dependence of the type described. Time, it should be noted, is always plotted as independent, when it constitutes one of the variables.

Functional Relationship

When the relationship between two variables is one of complete dependence, so that the value of y is uniquely determined by a given value of x, y is said to be a *function* of x. The general expression for such a relationship is $y = f(x)$. Thus the speed at which a body is falling at a

[1] It should be noted that letters at the end of the alphabet are used as symbols for *variables*, while letters at the beginning of the alphabet are used as symbols for *constants*, i.e., quantities the values of which do not change in the given discussion.

given moment is a function of the time it has been falling, the pressure of a given volume of gas is a function of its temperature, the increase of a given principal sum of money at a fixed rate of interest is a function of time. If the values of the independent variable be laid off on the x-axis of a rectilinear chart and the corresponding values of the function (i.e., the dependent variable) be laid off on the y-axis, a graphic representation of the function will be secured, in the form of a curve.[1] This concept of functional relationship is a very important one in statistical work. Some of the simpler functions may be briefly discussed.

THE STRAIGHT LINE

If two variables are so related that their values are always the same, their relationship is obviously of the form $y = x$. As a very simple example, the relation between the age of a tree and the number of rings in its trunk may be cited. A tree 6 years old will have 6 rings, for 20 years there will be 20 rings, and so on. This relationship may be represented on a coördinate chart, several sample values of x and y being taken. When these points are plotted and a line drawn through them, we secure a straight line passing through the origin and (assuming the two scales to be equal) bisecting the right angle XOY (cf. Fig. 3).

Similarly, any equation of the first degree (i.e., not involving xy, or powers of x or y other than the first) may be represented by a straight line. The generalized equation can be reduced to the form $y = a + bx$, where a is a constant representing the distance from the origin to the point of intersection of the given line and the y-axis, and b is a constant representing the slope of the given line (that is, the tangent of the angle which the line makes with the horizontal). The constant term a is called the *y-intercept*. It is clear from the generalized equation of the straight line that

[1] The general term "curve" is used to designate any line, straight or curved, when located with reference to a coördinate system.

when x has a value of zero, y will be equal to this constant term. In the example given above (Fig. 3) a is equal to 0, and b to 1. The location of a given line depends upon the signs of a and b as well as upon their magnitudes. The practical problem involved in the determination of any straight

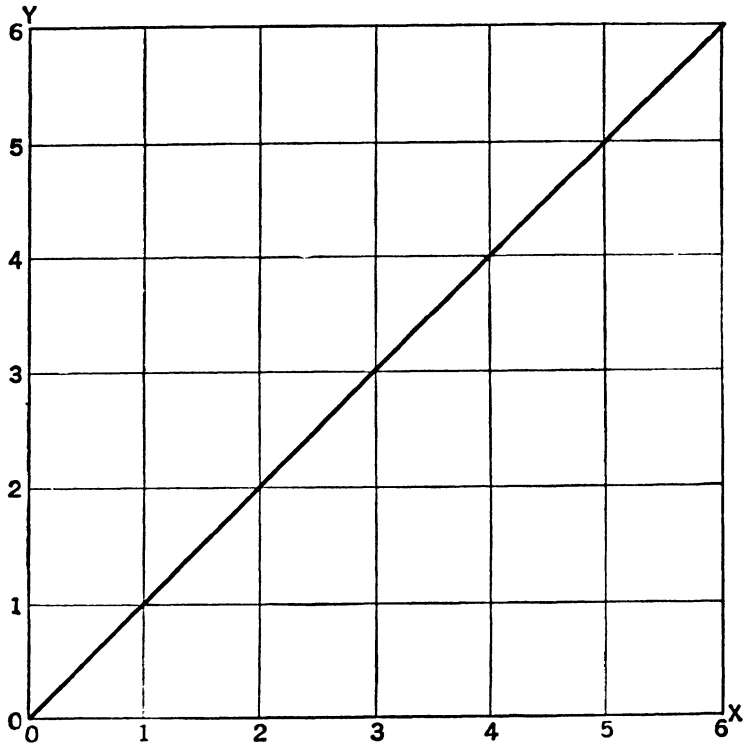


FIG. 3. — Graph of the Equation $y = x$

line is that of finding the values of a and b from the data, a problem which will appear in various forms in the discussion of statistical methods.

These points may be illustrated by the plotting of a simple equation of the first degree. Thus, to construct the graph of the function, $y = 2 + 3x$, various values of x are assumed, and corresponding values of y are determined. These may be arranged in the form of a table:

x	y $(2 + 3x)$
− 4	− 10
− 2	− 4
0	2
2	8
4	14

Plotting these values and connecting the plotted points, the graph illustrated in Fig. 4 is secured. It will be noted

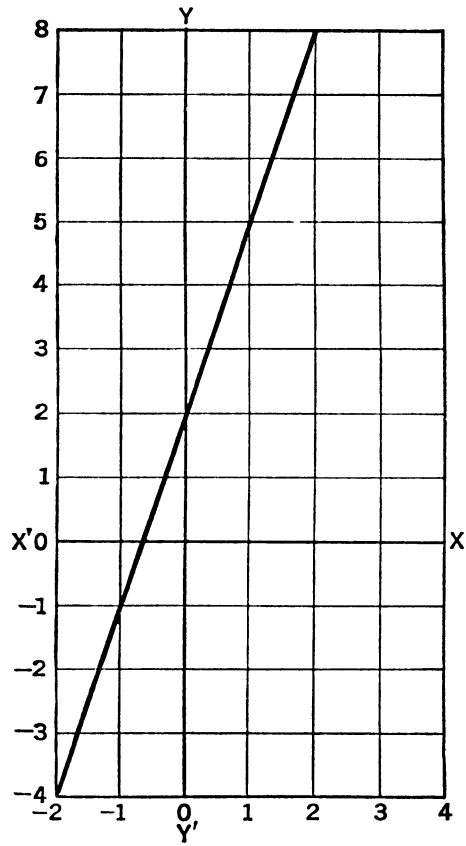


FIG. 4. — Graph of the Equation $y = 2 + 3x$

that since this function is linear (that is, the graph takes the form of a straight line) any two of the points would have

been sufficient to locate the line. The y-intercept is equal to the constant term 2, and the tangent of the angle which the given line makes with the horizontal (the slope of the line) is equal to 3, the coefficient of x. That this curve represents the equation is proved by the fact that the equation is satisfied by the coördinates of every point on the curve, and that every pair of values satisfying the equation is represented by a point on the curve. It is characteristic of a linear relationship that if one variable be increased by a constant amount, the corresponding increment of the other variable will be constant. In the above case as x grows by constant increments of 2, for example, the constant increment of the y-variable is 6. Series which increase in this way by constant increments are termed *arithmetic series*.

Many examples of linear relationship between variables are found in the physical sciences. An example from the economic world is found in the growth of money at simple interest, that is, interest which is not compounded. If we let r represent the rate of simple interest, x the number of years, and y the sum to which one dollar will amount at the end of x years, the equation of relationship is of the form

$$y = 1 + rx.$$

Since in a given case r will be constant, this is of the simple linear type. In statistical work precise relationships of this type rarely if ever occur, but approximations to the straight line relationship are found constantly.

NON-LINEAR RELATIONSHIP

Non-linear functions are of many types, of which only a few of the more common will be discussed here. The student should be familiar with the general characteristics of the chief non-periodic curves, of which the parabolic and hyperbolic types, on the one hand, and the exponential type on the other, are the most important. The potential

series is mentioned as a more general form of rather wide utility. Of periodic functions the sine curve is briefly described, as a fundamental form.

Functional relationships of the parabolic or hyperbolic form are quite common in the physical sciences, and such curves are found to fit certain classes of economic data. The general equation, when there is no constant term, is of the form $y = ax^b$. The curve is *parabolic* when the exponent b is positive, and *hyperbolic* when b is negative. The two following examples will serve to illustrate these types:

Problem: To construct the graph of the function $y = x^2$.

x	y (x^2)
-5	25
-4	16
-3	9
-2	4
-1	1
0	0
1	1
2	4
3	9
4	16
5	25

The graph is shown in Fig. 5.

Problem: To construct the graph of the function $y = x^{-1}$, for positive values of x.

x	y (x^{-1})
$\frac{1}{3}$	3
$\frac{1}{2}$	2
1	1
2	$\frac{1}{2}$
3	$\frac{1}{3}$
4	$\frac{1}{4}$
5	$\frac{1}{5}$

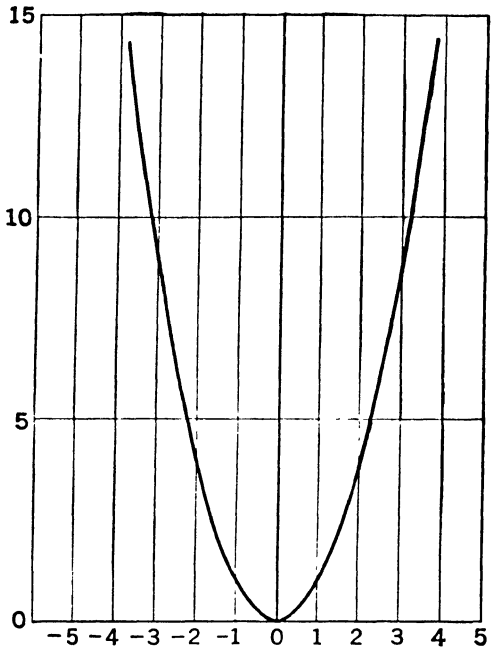


FIG. 5. — Parabola: Graph of the Equation $y = x^2$

The graph of this function, an equilateral hyperbola, is shown in Fig. 6. It should be noted that this equation may also be written $y = \frac{1}{x}$ or $xy = 1$.

It is characteristic of relationships of this type that as x increases in geometric progression, y also increases in geometric progression. Thus, in the example of the parabola given above ($y = x^2$), if we select the x values which form a geometric series,[1] the corresponding y values form a similar series:

x	1	2	4	8	16	32
y	1	4	16	64	256	1,024

Another class of functions is of the form represented by the equation $y = ab^x$. In equations of this type one of the variable quantities occurs as an exponent; graphs repre-

[1] A geometric series is one each term of which is derived from the preceding term by the application of a constant multiplier.

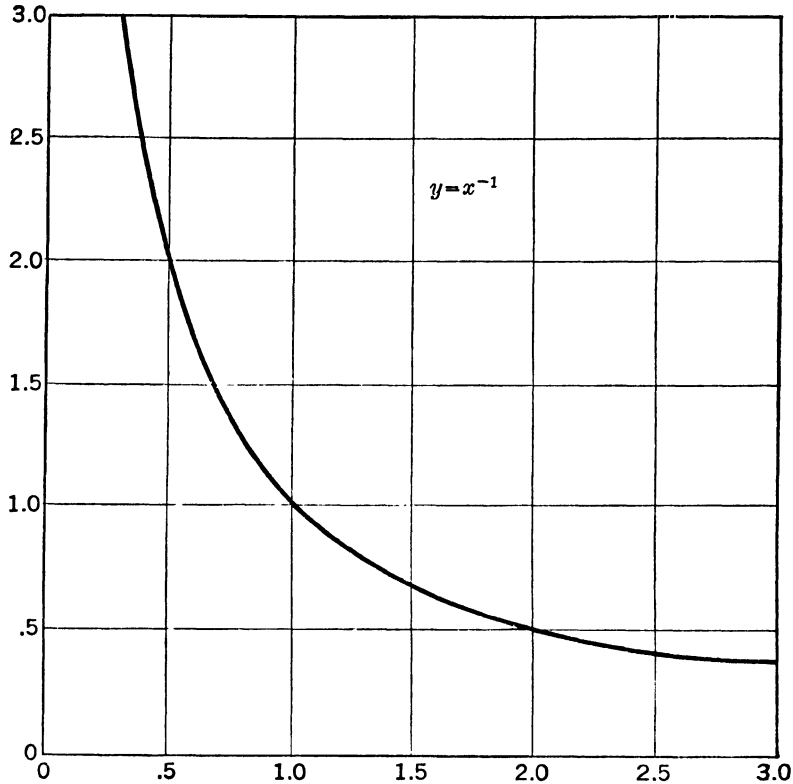


FIG. 6. — Equilateral Hyperbola: Graph of the Equation $y = x^{-1}$ (for positive values of x)

senting such equations are called *exponential curves.* The example which follows illustrates the type.

Problem: To construct the graph of the function $y = 2^x$, for positive values of x.

x	y (2^x)
0	1
1	2
2	4
3	8
4	16
5	32
6	64

This graph is shown in Fig. 7.

It has been noted that the relationship between two variables which increase by constant increments (constituting arithmetic series) may be represented by a straight line, and that the relationship between variables increasing

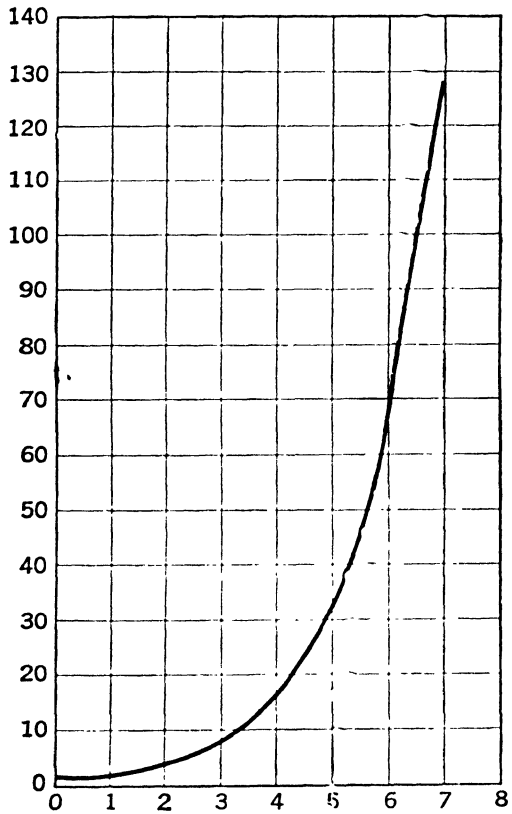


FIG. 7. — Exponential Curve: Graph of the Equation $y = 2^x$ (for positive values of x)

in geometric progression may be represented by either a parabola or a hyperbola. The exponential curve constitutes a hybrid type. It describes a relation in which one variable increases in arithmetic progression while the other increases in geometric progression. The figures given above illustrate this relationship.

Curves based upon relationships of the following type have been employed extensively in statistical inquiries:

$$y=a+bx+cx^2+dx^3+ \ldots .$$

The term *potential series* has been applied to equations of this type. Though such curves do not constitute parabolas of the strict conic section type, a curve based upon such an equation carried to the second power of x is termed a second degree parabola, to the third power of x, a third degree parabola, etc. No uniform and simple type is secured from this series. It is treated in more detail at a later point.

Periodic functions constitute another distinct type, a class represented notably by electrical and meteorological relations, though not confined to these fields. The characteristic feature of such relationships is that values of the dependent variable repeat themselves at constant intervals of the independent variable. The sine curve, the basic type of this class, is illustrated in the following example.

Problem: To construct the graph of the function $y = \sin x$.

x (angle in degrees)	y ($\sin x$)
0°	.000
30°	.500
60°	.866
90°	1.000
120°	.866
150°	.500
180°	.000
210°	− .500
240°	− .866
270°	− 1.000
300°	− .866
330°	− .500
360°	.000
390°	.500
etc.	

The graph is shown in Fig. 8.

The full importance in statistical work of securing a mathematical expression for the relation between two variables cannot be demonstrated until the subject has been further developed. One fundamental object is the determination of physical or economic laws underlying observed phenomena. Another more practical object is the securing

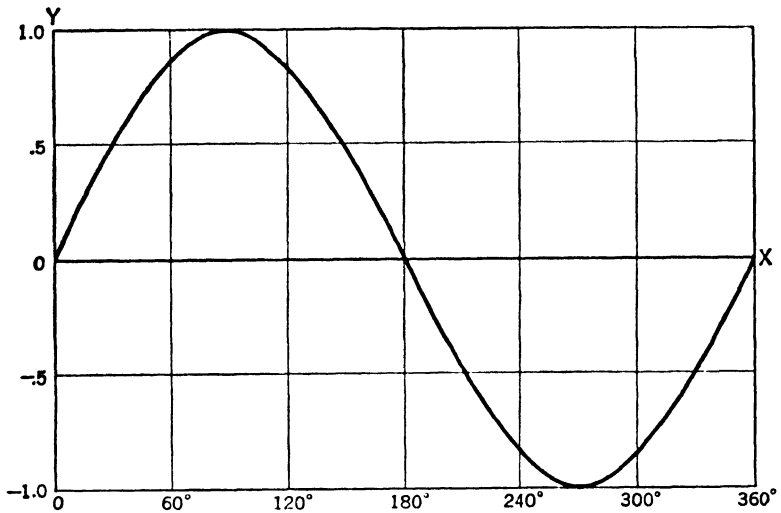


FIG. 8. — Sine Curve: Graph of the Equation $y = \text{Sin } x$

of a formula by means of which values of one variable may be approximated from given values of the other. Examples throughout the book will serve to illustrate how these objects are attained.[1]

LOGARITHMS

Logarithms, which play such an important part in general mathematical operations, are of equal importance in the manipulation of the raw materials of statistics. The nature of logarithms, and the methods by which they are employed to facilitate arithmetic processes, may be briefly

[1] A fuller discussion of different curve types is presented below, in the section dealing with the analysis of time series.

reviewed. This discussion is concerned only with the common system of logarithms of which the base is 10.

Any positive number may be expressed as a power of 10. Thus

$$1{,}000 = 10 \times 10 \times 10 = 10^3$$
$$10{,}000 = 10 \times 10 \times 10 \times 10 = 10^4.$$

In each case the *exponent* of 10 (the small number written above and to the right) indicates the number of times the figure 10 is repeated as a factor. For the integral powers of 10 the exponent is a whole number, but for the other numbers the exponent will contain a fractional value. Thus 100 is equal to 10 raised to the power 2, or 10^2; 110 is equal to 10 raised to the power 2.04139, or $10^{2.04139}$.

The exponent of 10, or the index of the power to which 10 must be raised to equal a certain number, is called the *logarithm* of that number. The logarithm of 100 is 2, the logarithm of 110 is 2.04139, the logarithm of 998 is 2.99913. These figures all have reference to the base 10, though a system of logarithms might be developed on any base. In general, if

$$a = b^c$$
$$\log_b a = c$$

which may be read "the logarithm of a to the base b is equal to c." The relation between the given number, the base and the logarithm, when the common system of logarithms is employed, may be easily remembered if the following relations are kept in mind:

$$100 = 10^2$$
$$\log_{10} 100 = 2.$$

The logarithm of any number has two parts, the integral and the decimal. The whole number is called the *characteristic*, and the decimal portion is termed the *mantissa*. The former is determined in a given case by inspection, while the mantissa may be obtained from logarithmic tables.

The characteristic varies with the location of the decimal point, while the mantissa remains the same for any given combination of numbers. This fact is illustrated by the following figures:

log of	8,450	= 3.92686
log of	845	= 2.92686
log of	84.5	= 1.92686
log of	8.45	= 0.92686
log of	0.845	= 9.92686 − 10
log of	0.0845	= 8.92686 − 10.

In finding the natural number to which a given logarithm corresponds (such natural numbers are termed *anti-logarithms*), the mantissa determines the sequence of figures, while the whole number, or characteristic, determines the location of the decimal point. For example, in seeking the anti-logarithm of 2.17609 it is found that the decimal .17609 follows the natural number 1,500, in a table of logarithms. Since the characteristic is 2, the natural number desired must lie between 100 and 1,000, and must therefore be 150.

A brief study of the following figures, showing the progression of numbers corresponding to certain powers of 10, will help to fix in mind the relations between the multiples of 10 and their logarithms, and will enable the characteristic of a desired logarithm to be readily determined.

.0001	.001	.01	.1	1	10	100	1,000	10,000
10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}	10^{3}	10^{4}

The exponents of 10 in the lower row are the logarithms of the numbers in the upper row.

It should be noted that the logarithms of all numbers from 0 to 1 are negative. Thus the logarithm of 0.845 is $-1 + .92686$; this is written 9.92686 − 10. In covering the range of all positive natural numbers from zero to infinity, logarithms traverse all positive and negative values.

A negative natural number, therefore, can have neither a positive nor a negative logarithm.

The advantage of thus expressing numbers as powers of 10 lies in the fact that the ordinary arithmetic operations of multiplication, division, raising to powers and extracting roots are greatly facilitated by this procedure.

To multiply numbers, add their logarithms. The sum of the logarithms of the factors is the logarithm of their product. In general terms:

$$a^b \times a^c = a^{(b+c)}.$$

Specifically, putting $a = 10$, $b = 2$, $c = 3$:

$$10^2 \times 10^3 = (10 \times 10) \times (10 \times 10 \times 10) = 10^5 = 100{,}000$$
$$100 \times 1{,}000 = 100{,}000.$$

To divide one number by another, subtract the logarithm of the latter from the logarithm of the former. The remainder is the logarithm of the desired quotient.

In general terms:

$$a^b \div a^c = a^{(b-c)}.$$

Specifically, putting $a = 10$, $b = 5$, $c = 2$:

$$10^5 \div 10^2 = \frac{10 \times 10 \times 10 \times 10 \times 10}{10 \times 10} = 10^3 = 1{,}000$$
$$100{,}000 \div 100 = 1{,}000.$$

To raise a given number to any power, multiply the logarithm of the number by the index of the power. The product is the logarithm of the desired power.

In general terms:

$$(a^b)^c = a^{bc}.$$

Specifically, putting $a = 10$, $b = 3$, $c = 2$:

$$(10^3)^2 = (10 \times 10 \times 10) \times (10 \times 10 \times 10) = 10^6 = 1{,}000{,}000$$
$$1{,}000^2 = 1{,}000{,}000.$$

To extract any root of a given number, divide the logarithm of the number by the index of the root. The quotient is the logarithm of the desired root.

In general terms:

$$\sqrt[b]{a^c} = a^{\left(\frac{c}{b}\right)}.$$

Specifically, putting $a = 10$, $b = 3$, $c = 6$:

$$\sqrt[3]{10^6} = 10^{\frac{6}{3}} = 10^2 = 100$$
$$\sqrt[3]{1,000,000} = 100.$$

In summary:

$$\log (a \times b) = \log a + \log b$$
$$\log (a \div b) = \log a - \log b$$
$$\log a^b = b \times \log a$$
$$\log \sqrt[b]{a} = \log a \div b.$$

These characteristic advantages of logarithms have been made use of in the construction of the slide rule, an instrument for reducing routine toil which should be familiar to all students of statistics.

LOGARITHMIC EQUATIONS

The graphic representation of data by means of a system of rectangular coördinates has been described above and some of the advantages of this method have been outlined. For many purposes it is desirable to plot logarithms rather than the natural numbers themselves. This may result in bringing out significant relations more distinctly, or it may serve greatly to simplify and facilitate the manipulation of data. In particular, when it is possible through the use of logarithms to reduce a complex curve to the straight line form, a distinct gain has been made in the direction of simplicity of treatment and interpretation.

A linear equation, it will be recalled, is of the general form $y = a + bx$, where a and b are constants which measure, respectively, the y-intercept of the given line and the slope. The simplification of equations through the use of logarithms involves in all cases the substitution of $\log x$ or $\log y$, or both, for the x or y variables, thereby reducing an equation of a higher order to a simpler form.

This process may be illustrated with reference to the equation $y = x^2$. When plotted on rectangular coördinates this equation gives a curve of the parabolic type (cf. Fig. 5).

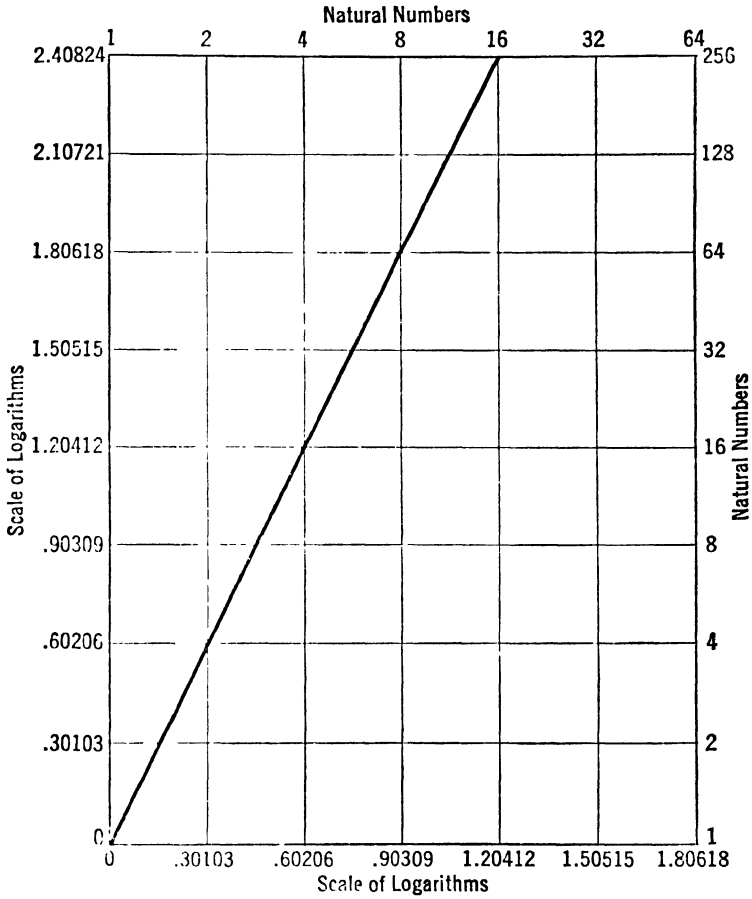


FIG. 9. Graph of the Equation $\log y = 2 \log x$ (Logarithmic form of the equation $y = x^2$)

Reduced to logarithmic form this becomes $\log y = 2 \log x$. This equation, in which the variables are $\log y$ and $\log x$, is linear in form. It is plotted in Fig. 9, for positive values of $\log x$. To indicate the relations involved, natural numbers corresponding to the logarithms are given on scales to the

right and at the top of the figure. The natural numbers appearing on the scales constitute geometric series, while their logarithms form arithmetic series. Equal vertical distances on the chart, it will be noted, represent equal absolute increments on the scale of logarithms and equal percentage increments on the scale of natural numbers.

The equation $y = 5x^3$ can be reduced in the same way to $\log y = \log 5 + 3 \log x$, a linear form. Similarly, all equations of the type $y = ax^b$, that is to say, all simple parabolas and hyperbolas, can be reduced to the straight line form $\log y = \log a + b \log x$. Graphically this means plotting the logarithms of the y's against the logarithms of the x's.

A different problem is presented by an equation of the type $y = ab^x$, the graph of which is termed an exponential curve. Expressed in logarithmic form, we have $\log y = \log a + x \log b$. This is also of the linear type, the two constants being $\log a$ and $\log b$, while the variables are x and $\log y$. If we plot the natural x's and the logs of the y's, with this type of equation, a straight line will be secured. A curve of this type is discussed and illustrated below.

LOGARITHMIC AND SEMI-LOGARITHMIC CHARTS

There are certain disadvantages to the plotting of logarithms, however. If a considerable number of points are being plotted the task of looking up the logarithms may be tedious, and, in addition, the original values, in which chief interest lies, will not appear on the chart. These difficulties may be avoided by constructing charts with the scales laid off logarithmically, but with the natural numbers instead of the logarithms appearing on the scales. This is an arrangement identical with that employed in the construction of slide rules. Thus, although the natural numbers are given on the scales, distances are proportional to the logarithms of the numbers thereon plotted. In Fig. 10

such a chart is presented, showing the graph of the equation $y = x^2$.

A variation of this type of chart which is of great importance in statistical work is one which is scaled arithmetically on the horizontal axis and logarithmically on the vertical axis. This is equivalent, of course, to plotting the

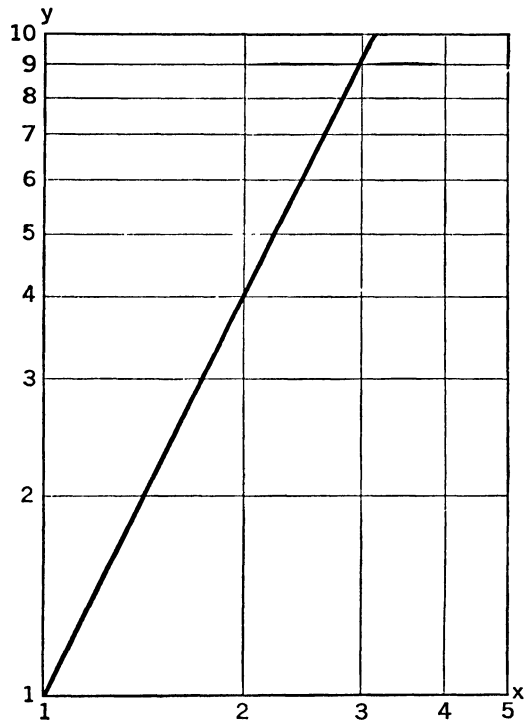


FIG. 10. — Graph of the Equation $y = x^2$ (Plotted on paper with logarithmic scales)

x's on the natural scale and plotting the logarithms of the y's. As was pointed out above, such a combination of scales reduces a curve of the exponential type to a straight line. Plotting paper of this semi-logarithmic or "ratio" type may be constructed with the aid of a slide rule or of logarithms, or may be purchased ready made. It is of

particular value in charting economic statistics, because of the fact that time is usually one of the variables in such cases, and it is desirable to plot this variable on the natural scale.

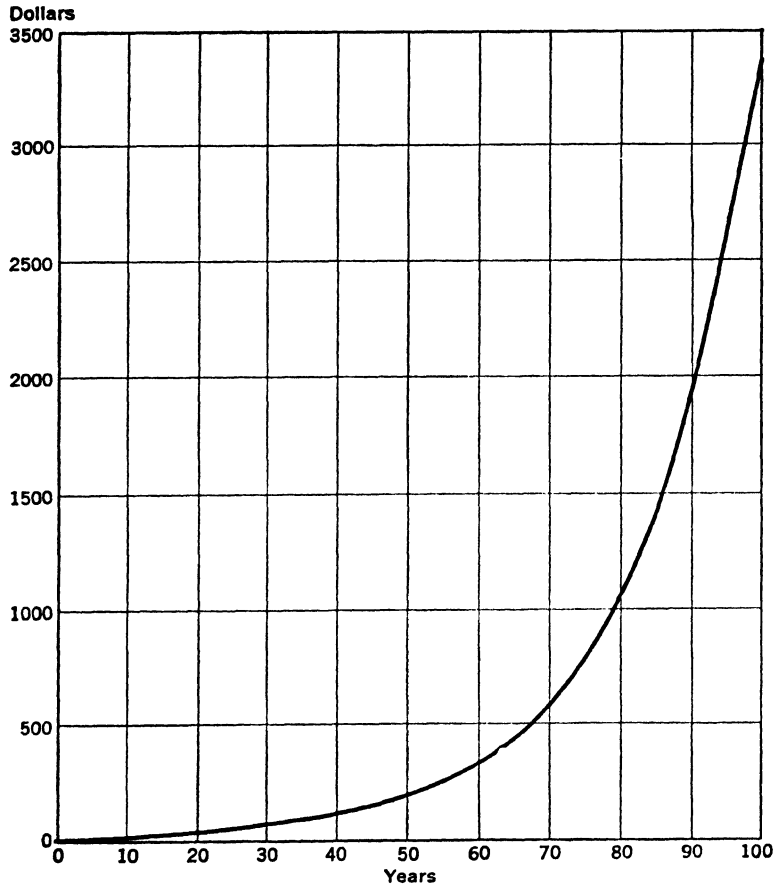


FIG. 11. — The Compound Interest Law: Growth of $10.00 at Compound Interest at 6 per cent for 100 Years (Plotted on arithmetic scale)

As an example of this type of curve the compound interest law may be used. If r be taken to represent the rate of interest, x the number of years, p the principal, and y the sum to which the principal amounts at the end of x

years (interest being compounded annually), an equation is secured of the form

$$y = p(1 + r)^x.$$

Expressed logarithmically this becomes

$$\log y = \log p + x \log (1 + r),$$

the equation to a straight line.

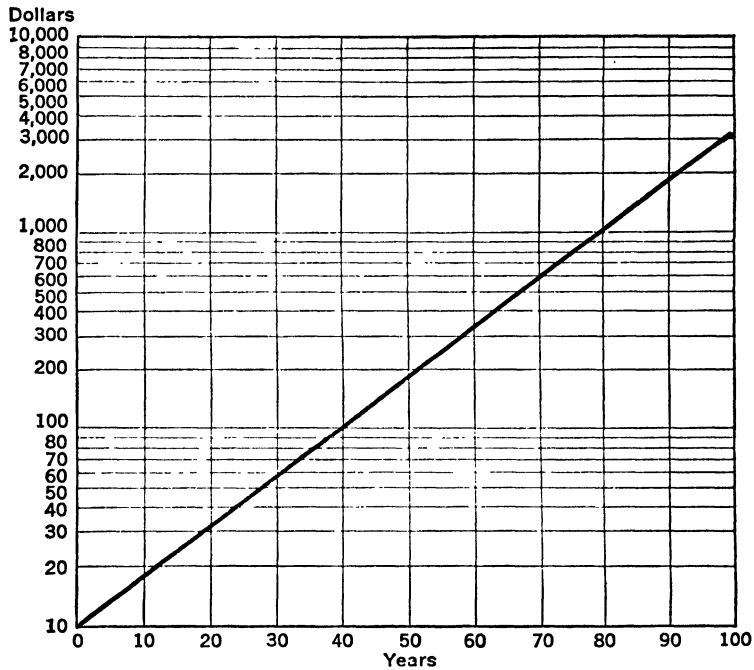


FIG. 12. — The Compound Interest Law: Growth of $10.00 at Compound Interest at 6 per cent for 100 Years (Plotted on semi-logarithmic or ratio scale)

In Fig. 11 a curve representing the growth of ten dollars at compound interest at 6 per cent is plotted on the natural scale. This is the graph of the exponential equation

$$y = 10(1 + .06)^x$$

y representing the total amount of principal and interest

at the end of x years. Figure 12 shows the same data plotted on semi-logarithmic paper, the exponential curve being reduced to a straight line.

The use of semi-logarithmic paper is not confined to cases in which an exponential curve is straightened out, for the significance of many types of data is most effectively brought out when charts of this type are used. These advantages are more fully explained below.

The Construction of Charts

When the results of observations or statistical investigations have been secured in quantitative form, one of the first steps toward analysis and interpretation of the data is that of presenting these results graphically. Not only is such procedure of scientific value in paving the way for further investigation of relationships, but it serves an immediate practical purpose in visualizing the results. A visual stimulus opens up a far more direct path to our understanding and imagination than that afforded by the more recently developed processes of reasoning. The interpretation of a column of raw figures may be a difficult task; the same data in graphic form may tell a simple and easily understood story. For these reasons graphic methods of presentation have come to play a highly important part in the everyday activities of business, as well as in the laboratory and drafting room.

It is beyond the scope of this book to present any detailed account of the multiplicity of graphs employed by engineers and statisticians today. Certain of the more important principles of graphic presentation may be briefly explained, however, and some of the chief types of graphs which are in daily use may be illustrated. Other examples appear in later chapters of this book.

Factors Governing the Selection of a Chart

The selection of the type of chart to be employed in a given case will depend upon two general considerations.

The first of these relates to the character of the material to be plotted. While the data of a given problem may frequently be presented graphically in several different forms, there is generally one type of chart best adapted to that material. It may be true, also, that certain types would be quite inappropriate to the data in question. The selection of a type of chart to employ, therefore, must be made with the characteristics of the data clearly in mind.

Perhaps more important is the *purpose* which the given chart is designed to serve. Each of the many types of charts in common use is appropriate to certain specific purposes. It will bring out certain characteristics of the data or will emphasize certain relationships. There is no chart which is sovereign for all purposes. Until the purpose is clearly defined the best chart form cannot be selected. The following descriptions of a few standard types will facilitate the selection of an appropriate form.

CHARTS ADAPTED TO THE PLOTTING OF TIME SERIES

In the graphic presentation of a time series, primary interest attaches to the chronological variations in the values of the data, to the general trend and to the fluctuations about the trend. If the purpose is to emphasize the *absolute* variations, the differences in absolute units between the values of the series at different times, a simple chart of the type illustrated in Fig. 13 will serve the purpose. This chart depicts annual wheat flour exports from the United States during the period 1913–1936. Both scales are arithmetic. Points representing the various annual values are shown and, to facilitate interpretation, these points are connected by a series of straight lines. The chart tells a simple story of year-to-year fluctuations, with a sharp advance at the end of the World War, a decline as the post-war emergency passed, several years of moderate growth, and a

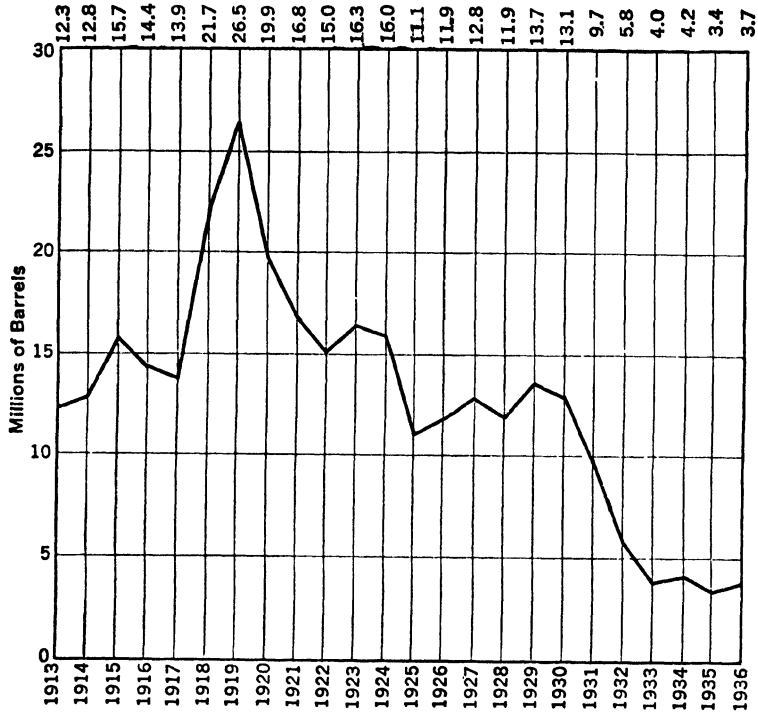


FIG. 13. — Wheat Flour Exports from the United States, 1913–1936

severe decline as the world depression deepened in the early thirties. With respect to general make-up, the following points should be noted:

1. The title constitutes a clear description of the material plotted and indicates the period covered.
2. The vertical scale begins at the zero line, enabling a true impression to be gained of the magnitude of the fluctuations.
3. The zero line and the line joining the plotted points are ruled more heavily than the coördinate lines.
4. Figures for the scales are placed at the left and at the bottom of the chart. The vertical scale may be repeated at the right to facilitate reading. All figures are so placed that they may be read from the base as bottom or from the right hand edge of the chart as bottom.

5. The y-values of the plotted points are given at the top of the chart. This practice is helpful, though not necessary, as the values may be presented in a separate table.

ADVANTAGES OF THE RATIO CHART

If *relative* rather than *absolute* variations are of chief concern, the chart employed should be of the semi-logarithmic type, scaled logarithmically on the y-axis and arithmetically on the x-axis. In such a chart equal percentage variations are represented by equal vertical distances, as opposed to the ordinary arithmetic type in which equal absolute variations are represented by equal vertical distances. The argument for the use of the semi-logarithmic or ratio chart for the representation of time series is that, in general, the significance of a given change depends upon the magnitude of the base from which the change is meas-

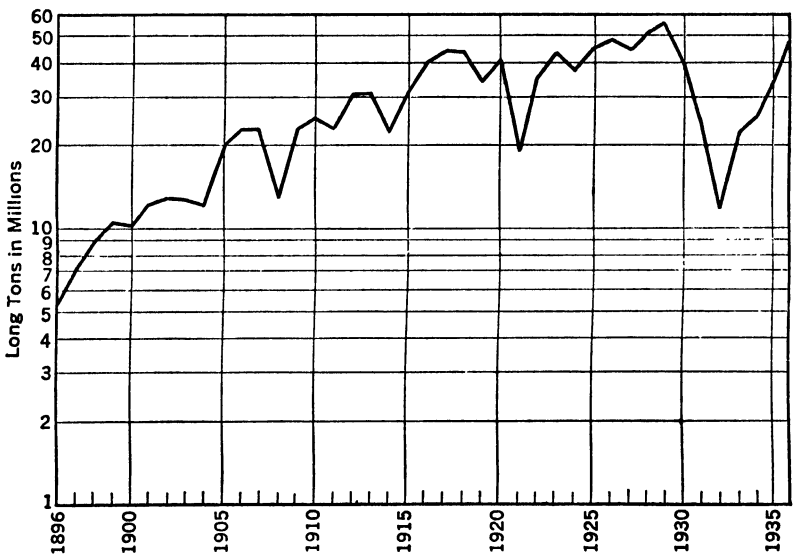


FIG. 14. — Production of Steel Ingots and Castings in the United States, 1896–1936 (Plotted on semi-logarithmic scale)

ured. That is, an increase of 100 on a base of 100 is as significant as an increase of 10,000 on a base of 10,000. In

each case there is an increase of 100 per cent. The absolute increase in the second case is 100 times that in the first case, and the two changes would show in this proportion on the arithmetic chart. They would show as of equal importance on the semi-logarithmic chart.

Such a chart is presented in Fig. 14, which shows the course of steel production in the United States from 1896 to 1936. The absolute magnitudes are plotted, but the vertical scale is so constructed as to represent variations from year to year in proportion to their relative magnitude.

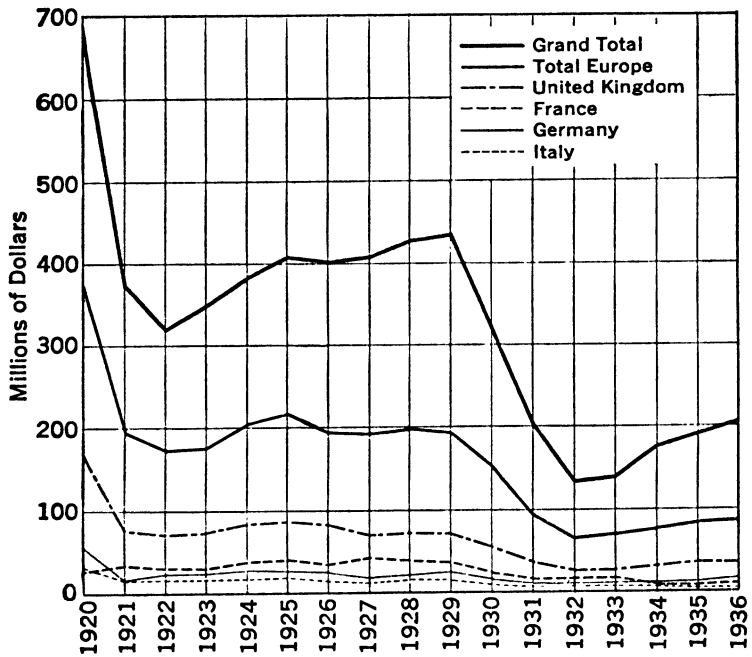


FIG. 15. — Exports of the United States, 1920–1936 Showing Total Exports and Exports to Selected Areas (Monthly averages for the years named are plotted on an arithmetic scale)

Certain distinctive advantages of the ratio or logarithmic ruling are brought out by a comparison of Fig. 15 and Fig. 16. The data presented graphically in these two charts are shown in Table 2:

TABLE 2

Exports of the United States, 1920–1936

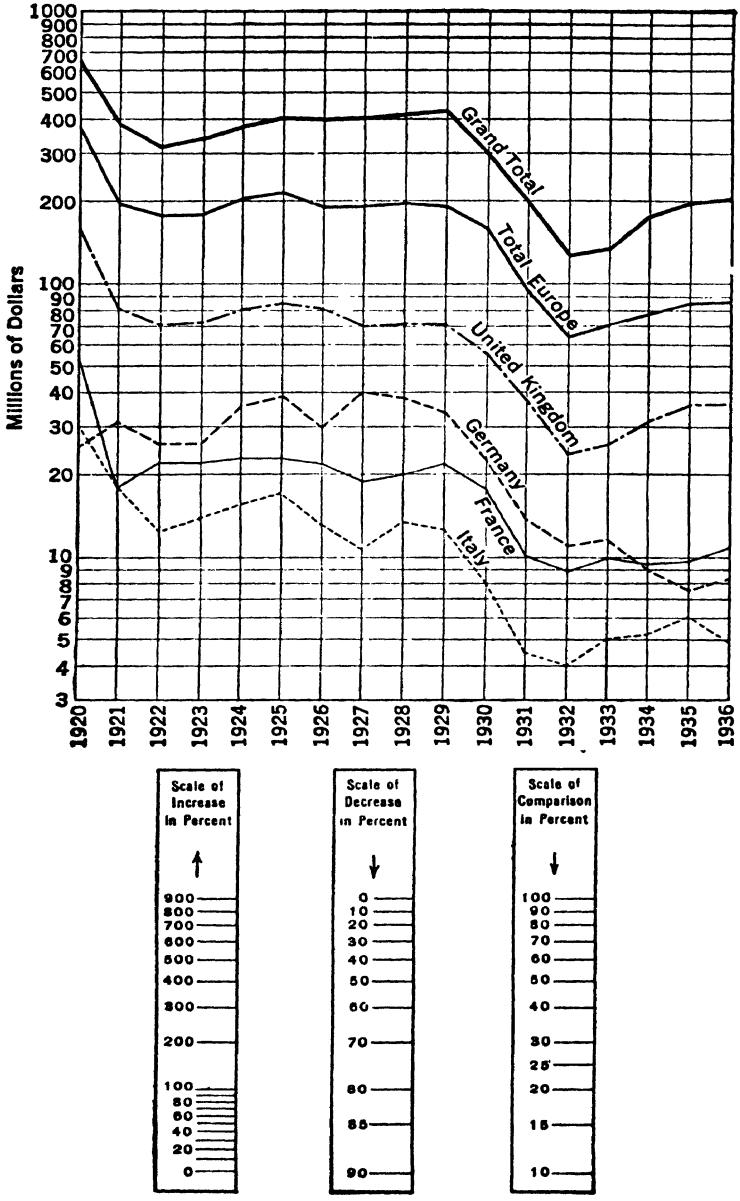
(Monthly averages, in thousands of dollars)

	To France	*To Germany*	*To Italy*	*To United Kingdom*	*Total to Europe*	*Grand total*
1920	$56,349	$25,953	$30,980	$161,319	$372,174	$685,668
1921	18,745	31,027	17,955	78,510	196,992	373,753
1922	22,247	26,343	12,575	71,319	173,613	319,315
1923	22,678	26,403	13,961	73,527	174,451	347,291
1924	23,472	36,702	15,595	81,912	203,775	382,582
1925	23,358	39,195	17,096	86,155	216,979	409,154
1926	22,000	30,347	13,117	81,051	192,512	400,722
1927	19,065	40,140	10,971	70,005	192,576	405,448
1928	20,058	38,938	13,510	70,613	197,912	427,363
1929	22,133	34,204	12,831	70,667	195,070	435,083
1930	18,663	23,189	8,369	56,509	153,198	320,265
1931	10,152	13,838	4,568	37,923	95,040	202,024
1932	9,297	11,139	4,095	24,027	65,358	134,251
1933	10,143	11,669	5,103	25,978	70,815	139,583
1934	9,642	9,062	5,381	31,896	79,150	177,733
1935	9,751	7,665	6,035	36,117	85,770	190,240
1936	10,795	8,382	4,900	36,662	86,694	204,457

(Data compiled by Bureau of Foreign and Domestic Commerce, U. S. Department of Commerce.)

If the six series are to be presented on a single chart, scaled arithmetically, a scale must be selected which will include the largest item recorded, $685,668,000. Such a scale reduces the relative importance of the smaller magnitudes. From Fig. 15 it appears that during the period covered by the chart very large fluctuations occurred in total exports, substantial but somewhat smaller movements occurred in exports to Europe, and that exports to the four individual countries suffered much less severe fluctuations. Such a picture is quite misleading. The true state of affairs is reflected in Fig. 16, in which the same data are plotted on paper with a semi-logarithmic ruling. Fluctuations in exports to the individual countries are here seen to have been relatively greater than the movements of total exports. For the purpose of comparing series which differ materially

FIG. 16. — Exports of the United States, 1920–1936, Showing Total Exports and Exports to Selected Areas, with Scales of Increase, Decrease, and Comparison (Monthly averages for the years named are plotted on a semi-logarithmic scale)



with respect to the magnitude of the individual items, the arithmetic ruling is quite useless, giving a thoroughly distorted picture of the true relations. The ratio ruling permits a legitimate comparison.

The scales printed below Fig. 16 emphasize certain very useful features of the logarithmic ruling. The *scale of increase* may be used to measure with a fair degree of accuracy the increase in a given series between any two dates. A given vertical distance on the chart, it will be recalled, represents a constant percentage increase at all points on the chart. Thus the distance from 1 to 10, along the vertical scale, is the same as the distance from 100 to 1,000. Any vertical distance may be measured, and the percentage of increase which it represents may be determined by laying off the given distance along the scale of increase, which is always read from the bottom up. For example, to determine the degree of increase in total exports from 1932 to 1935, we measure the vertical distance between the points plotted for these two years. Laying off this distance along the scale, it is found to represent about a 40 per cent increase.

The *scale of decrease* is used in a similar fashion. The vertical distance between any two points is measured, and the percentage decrease which it represents is determined by laying off the given distance on the scale from the top downward. The arrows indicate the direction in which the various scales are to be read.

By means of the *scale of comparison* the percentage relation of one series to another at any time may be determined. For example, we may wish to know the percentage relation between exports to Europe and total exports in 1935. The vertical distance between the two plotted points is measured, and laid off on the scale of comparison, reading from the top downward. It is found to be approximately 45 per cent.

Scales of the type illustrated above may be readily constructed on a given chart by using the ratio ruling for the scale intervals. When a series of charts is prepared on

semi-logarithmic paper of a standard type it is convenient to construct such scales in a more permanent form, in the shape of special rulers.

A ratio chart is particularly useful when interest attaches to rates of growth (or decline) over a considerable period of time. In such a case, the reading of the chart is facilitated by the plotting of straight diagonal lines indicating uniform

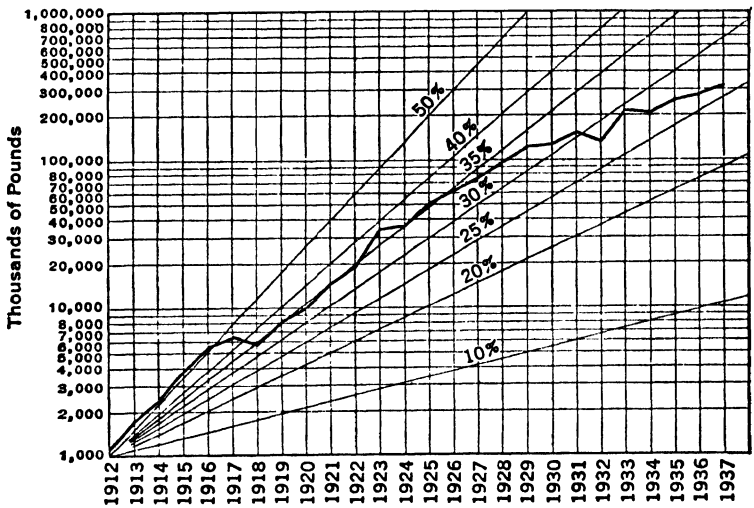


FIG. 17. — Production of Rayon Filament Yarn in the United States, 1912–1937, With Lines Defining Uniform Rates of Growth

rates of change. These should radiate from a single point of origin. The procedure is illustrated in Fig. 17. The diagonal lines there shown indicate changes at uniform rates ranging from 10 per cent to 50 per cent per year. By reference to these lines the user of the chart may readily determine the approximate rate of growth of the plotted series between any two years.

The chief advantages of the semi-logarithmic ruling in chart construction may be briefly summarized:

1. A curve of the exponential type becomes a straight line when plotted on a semi-logarithmic chart. For example, a curve

representing the growth of any sum of money at compound interest takes the form of a straight line when so plotted.

2. In any series, so long as the *rate* of increase or decrease remains constant the graph will be a straight line on this ruling.
3. Equal relative changes are represented by lines having equal slopes. Thus two series increasing or decreasing at equal rates will be represented by parallel lines.
4. Comparison of the rates of change in two or more series is effected by comparison of the slopes of the plotted lines.
5. The semi-logarithmic ruling permits, at the same time, the plotting of absolute magnitudes and the comparison of relative changes.
6. Comparison of series differing materially in the magnitude of individual items is possible with the semi-logarithmic chart.
7. Percentages of change may be read and percentage relations between magnitudes determined directly from the chart.

CHARTS FOR THE COMPARISON OF FREQUENCIES

A different type of chart is called for when the object is the comparison of frequencies, that is, numbers of events or things of different classes. The following census figures may serve to illustrate the problem.

TABLE 3

Farms in New England States in 1935

State	*Number of farms*
Maine	41,907
New Hampshire	17,695
Vermont	27,061
Massachusetts	35,094
Rhode Island	4,327
Connecticut	32,157

A graphic comparison of these six states with respect to number of farms in 1935 is afforded by the bar diagram in Fig. 18. This is a simple but effective type of chart for this purpose.

Further examples of this type of chart, as employed in the representation of frequency distributions, are contained in the next chapter. It is there shown how a frequency polygon or frequency curve may grow out of the simple bar diagram, when data of certain kinds are being handled.

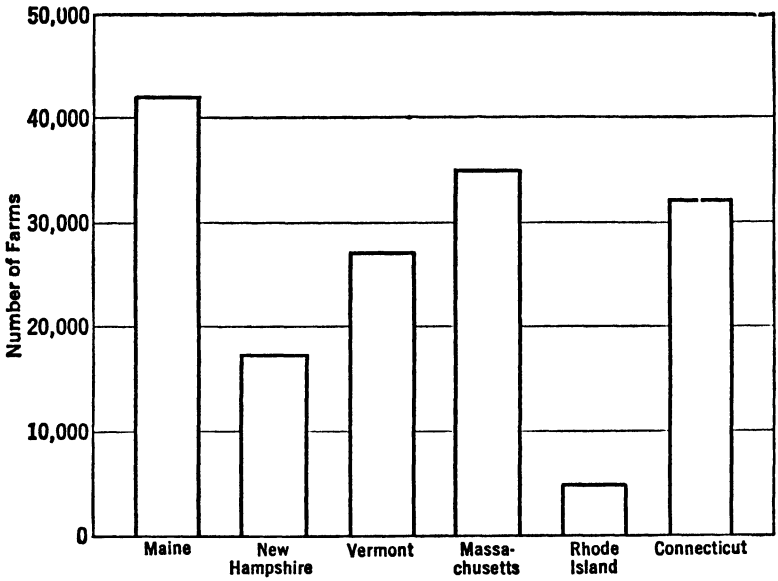


FIG. 18. — Farms in New England States in 1935

Such frequency curves constitute very important graphic types, but it will be more appropriate to treat them in full at a later point.

CHARTS FOR THE REPRESENTATION OF COMPONENT PARTS

It is frequently desirable in tabular and graphic presentation to break up a total into its component parts, in order that changes in the parts as well as in the total may be followed. The table on page 43 exemplifies this procedure.

These figures are presented graphically in Fig. 19, which reveals the varying post-war fortunes of different interests in American manufacturing industries. It is clear from the diagram that the general swings of material costs, labor costs and overhead costs in American manufacturing industries have paralleled the fluctuations in total value of products. Some of the movements of the component items are of exceptional interest, however. Overhead costs (with which

TABLE 4

Total Value of Products and Elements of Production Costs, Manufacturing Industries of the United States, 1919–1935

(Millions of dollars)

Year	Cost of materials [1]	Labor cost (wages)	Overhead cost plus profits [2]	Total value of products
1919	$37,233	$10,462	$14,347	$62,042
1921	25,321	8,202	10,130	43,653
1923	34,706	11,009	14,841	60,556
1925	35,936	10,730	16,048	62,714
1927	34,803	10,836	16,639	62,278
1929	38,178	11,607	20,176	69,961
1931	21,681	7,173	12,184	41,038
1933	16,821	5,262	9,276	31,359
1935	26,264	7,545	11,951	45,760

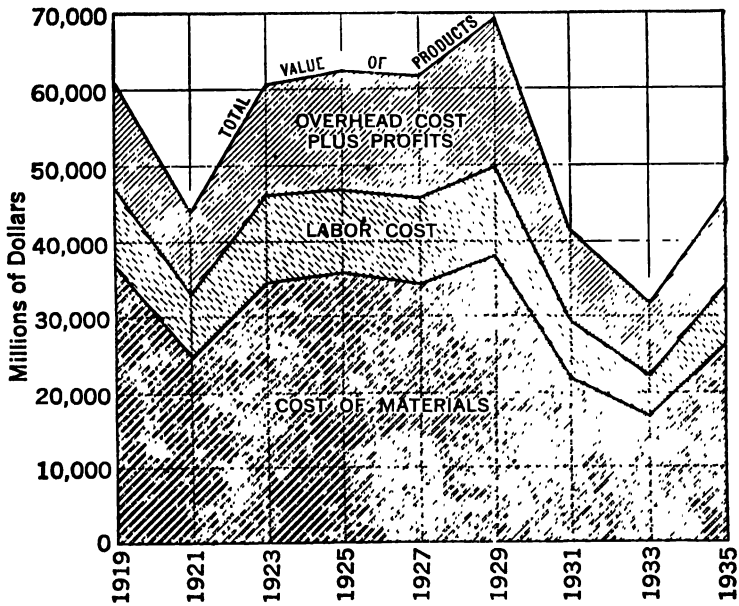


FIG. 19. — Total Value of Products and Elements of Production Costs, Manufacturing Industries of the United States, 1919–1935

[1] Including containers, fuel and purchased electric energy.

[2] This item represents the difference between total direct costs (materials and wages) and total value of products. It includes overhead costs proper, plus salaries, taxes, profits, etc.

profits are here combined) showed a notable expansion between 1921 and 1929. The great recession that followed squeezed all the elements of the total, forcing them to levels well below those of the 1921 depression.

CUMULATIVE CHARTS

In many cases chief interest in the development of a series attaches not to the value of each successive item but to the cumulated total of a number of such items. This may be so when a yearly production program has been laid out. In such a case it is the relation between cumulated production to date and scheduled production to date which is of major interest, and a chart form is needed which will enable this comparison to be made. The following figures illustrate the type of data for which such charts are appropriate.

TABLE 5

Cumulative Production Schedule and Cumulative Output, 1936

(Speedwell Automobile Company)

Month	*Production schedule (cars)*	*Cumulative production schedule (cars)*	*Output (cars)*	*Cumulative output (cars)*
January	8,000	8,000	6,125	6,125
February	10,000	18,000	9,250	15,375
March	12,000	30,000	10,514	25,889
April	15,000	45,000	15,131	41,020
May	14,000	59,000	12,159	53,179
June	12,000	71,000	13,250	66,429
July	11,000	82,000	11,462	77,891
August	10,000	92,000	10,531	88,422
September	6,000	98,000	4,621	93,043
October	9,000	107,000	9,843	102,886
November	10,000	117,000	13,785	116,671
December	10,000	127,000		

It is assumed that this table represents the situation as of the end of November.

In Fig. 20 the two cumulative curves are plotted. The relation between actual and scheduled production at the end of each month is shown on the chart, and it is possible

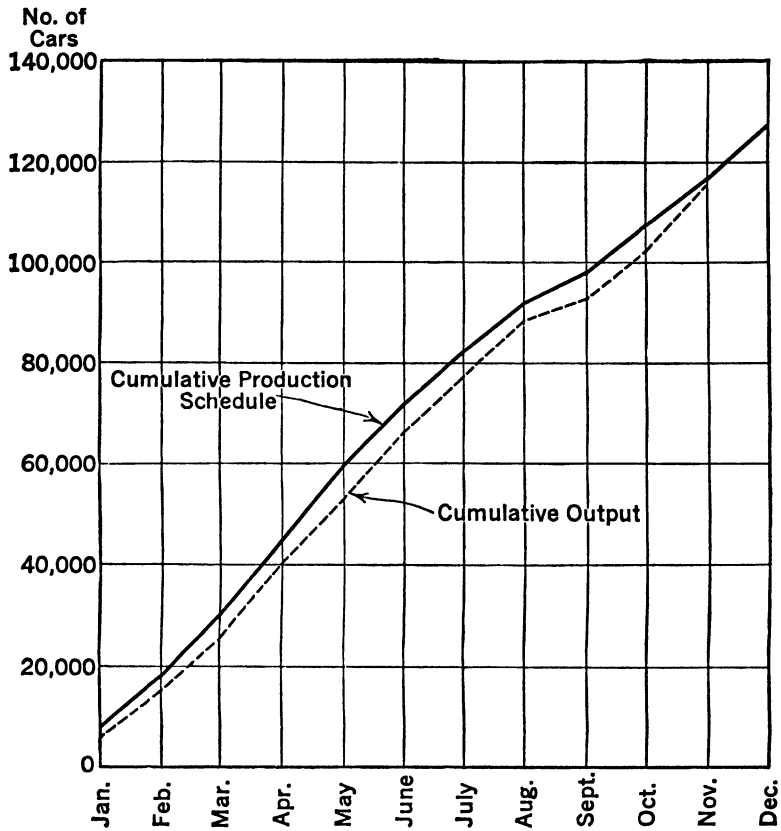


FIG. 20. — Comparison of Scheduled and Actual Output (Cumulative) Speedwell Automobile Co. 1936

from the scale to read the approximate amount by which production is behind schedule. By reference to the figures, which should always accompany the chart, the exact relation may be determined. Such a chart has many applications, some of which are illustrated in the following chapter.

THE GANTT PROGRESS CHART

The same data may be presented in a very effective form by making use of a type of chart developed by Mr. H. L. Gantt. An adequate description of this chart and of its many uses would far exceed the space which can be given to it here, but its characteristics may be indicated in a very brief account.

Once a schedule has been drawn up, the Gantt chart may be utilized in checking actual accomplishment against the schedule. Having such a schedule as that given in Table 5, the monthly and annual quotas may be entered on a form similar to that shown in Fig. 21. The entry to the left of each monthly space indicates the amount scheduled for production during that month. The entry to the right of each monthly space indicates the cumulated scheduled production to the end of the given month. In this figure the results of the first two months' operations are shown. The heavy black line indicates the cumulated actual production during this period, amounting to 15,375 cars. The narrow upper lines in the January and February columns measure the actual production in each of those months. If actual production in either month had equaled the scheduled production the light line would extend across the full monthly space. When actual production in a given month exceeds the scheduled production a double light line appears.

It should be noted that the spaces into which each monthly period is divided represent equal parts of a month but varying amounts in terms of actual production. Thus the space representing one fifth of the January interval represents a production of 1,600 cars (the January quota being 8,000). The space representing one fifth of the April interval represents 3,000 cars (the April quota being 15,000). In reading the chart in terms of absolute magnitudes reference must be had to the monthly quotas.

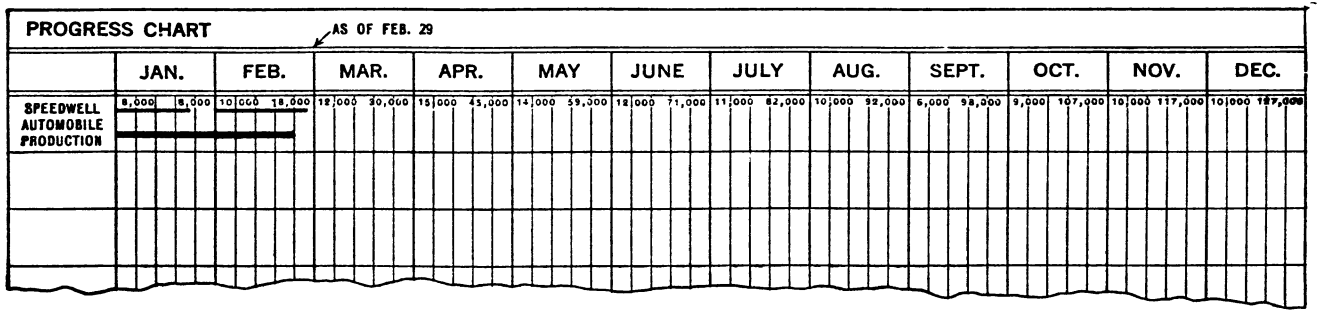


FIG. 21. — Comparison of Scheduled and Actual Output, 1936: Gantt Progress Chart (Showing the situation on February 29th)

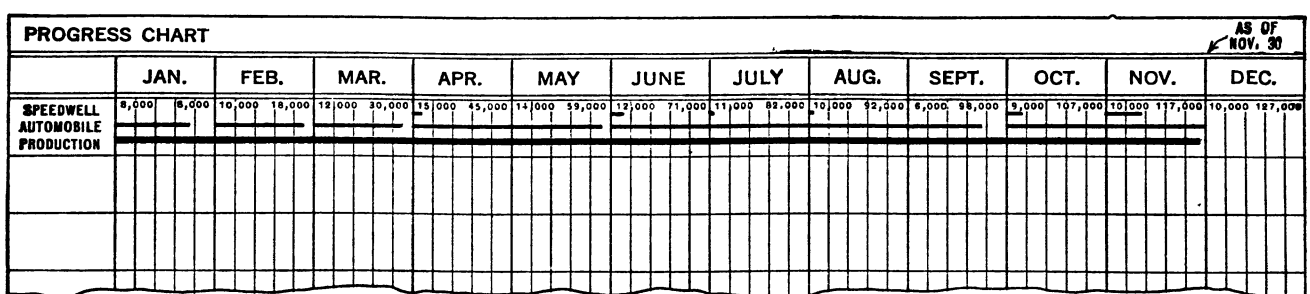


FIG. 22. — Comparison of Scheduled and Actual Output, 1936: Gantt Progress Chart (Showing the situation on November 30th)

The situation at the end of November is shown in Fig. 22. The arrow at the top of the diagram indicates the point of time actually reached. That actual production is slightly behind scheduled production is apparent from the relation between this arrow and the heavy black line, while the light lines indicating monthly production show that actual output has exceeded the monthly quota in five of the last six months.

The Gantt chart has a great variety of applications in governmental and business organizations. The economy of space is such that developments in a number of departments or districts may be shown on a single chart. It constitutes the simplest and most effective graphic method known for following the progress of work under way, for comparing actual accomplishment with an established program. And in so doing, it increases by so much the efficiency of administrative control.

Preferred Practice for Graphic Presentation

Graphic methods have been widely employed in the physical and social sciences and in business, and the resulting diversity of uses has made it difficult to secure standardization of practice. To remedy this defect a committee representing engineering, statistical and research organizations was organized in 1929, under the sponsorship of the American Society of Mechanical Engineers, for the purpose of formulating principles of preferred practice in this field. This group, acting as a Sectional Committee of the American Standards Association, is compiling a code of preferred practice for graphic presentation. The first section of this code, dealing with time series charts, has been issued by the sectional committee. This report furnishes an excellent summary of conventional procedures, with detailed recommendations concerning principles appropriate to the graphic presentation of time series. Somewhat more specialized, although it deals with certain principles applicable to the entire field of

graphics, is another report of the same committee on charts suitable for use as lantern slides.[1]

REFERENCES

American Recommended Practice, Engineering and Scientific Charts for Lantern Slides. American Society of Mechanical Engineers.

Arkin, H. and Colton, R. R. *Graphs: How to Make and Use Them.*

Brinton, W. C. *Graphic Methods for Presenting Facts.*

Clark, Wallace. *The Gantt Chart.*

Code of Preferred Practice for Graphic Presentation: Time Series Charts. American Society of Mechanical Engineers.

Griffin, F. L. *Introduction to Mathematical Analysis.*

Haskell, A. C. *How to Make and Use Graphic Charts.*

Karsten, Karl G. *Charts and Graphs.*

Lipka, Joseph. *Graphical and Mechanical Computation.*

Mudgett, Bruce D. *Statistical Tables and Graphs.*

Riggleman, John R. *Graphic Methods for Presenting Business Statistics.*

Schultze, Arthur. *Graphic Algebra.*

Steinmetz, C. P. *Engineering Mathematics.*

(The publishers and the dates of publication of the volumes named above are given in the bibliography at the end of this volume.)

[1] The titles of these reports are, respectively:

Code of Preferred Practice for Graphic Presentation: Time Series Charts. Prepared by Subcommittee on Preferred Practice in Graphic Presentation, A. H. Richardson, Chairman, of the Sectional Committee on Standards for Graphic Presentation under Procedure of American Standards Association. Sponsored by the American Society of Mechanical Engineers, 29 West 39th Street, New York. 68 pages. (Preliminary Report.)

American Recommended Practice, Engineering and Scientific Charts for Lantern Slides. Prepared by Subcommittee on Engineering and Scientific Charts, Walter A. Shewhart, Chairman, of the Sectional Committee on Standards for Graphic Presentation. American Society of Mechanical Engineers, New York. 10 pages.

CHAPTER III

THE ORGANIZATION OF STATISTICAL DATA: THE FREQUENCY DISTRIBUTION

The task of the statistician engaged in business or economic research includes the organization, analysis and interpretation of quantitative data relating to business affairs and to economic conditions. To these fundamental operations that of collecting the original data may be added, though more frequently data will be compiled directly from primary or secondary sources.

At the outset it is necessary to distinguish between the problems arising in the analysis of time series and those involved in the organization and analysis of materials in connection with which the time factor does not enter. In studying a time series the primary object is to measure and analyze the chronological variations in the value of the variable. Thus one may study variations in sales over a period of years, fluctuations in the production of bituminous coal, or changes in the general level of prices. Quite different is the procedure in the study of such a problem as income distribution at a given time. In this case we are desirous of knowing how many people in the United States fall in each of a number of income classes. The general problem of organization in this latter class of cases is to determine how many times each value of a variable is repeated and how these values are distributed. Data of this sort, when organized, constitute a *frequency series*, as opposed to the *time* or *historical series*. The methods appropriate to these two types of analysis differ fundamentally and will therefore be treated separately. In the present section we are concerned with the organization and preliminary analy-

sis of data in connection with which the time element, while it may be present, does not enter as a factor.

Unorganized Data

When quantitative data of the type with which the statistician works are presented in a raw state they appear as unorganized masses of material, without form or structure. They may have been drawn from the production or sales records of a business establishment, or they may represent a miscellaneous collection of price quotations. If the data have been gathered by other agencies they may already have been arranged in the form of a general table, but this form may be entirely unsuited to the particular object in the mind of the investigator. The first task of the statistician is the organization of the figures in such a form that their significance, for the purpose in hand, may be appreciated, that comparison with masses of similar data may be facilitated, and that further analysis may be possible. Scientific method, it has been noted, involves *observation*, *inference*, and *verification*. Data, the results of *observation*, must be put into definite form and given coherent structure before the process of *inference* is possible.

The figures on page 52, representing the earnings during a given week of 210 individuals engaged in piece work in a certain manufacturing establishment, will serve as an example of such data in their raw state.

The Array

If these figures are arranged in order of magnitude something will have been done toward securing a coherent structure. The range covered and the general distribution throughout this range will then be clear, and the way will be prepared for further organization. When so arranged the *array* on page 53 is secured.

WEEKLY EARNINGS OF 210 EMPLOYEES

$26.25	$28.70	$24.15	$29.75	$29.20	$30.60	$23.40	$24.75
26.70	24.35	25.75	27.20	28.30	25.25	27.75	27.60
28.20	27.30	27.80	26.35	27.40	28.30	26.60	25.75
27.70	28.60	25.30	27.80	26.40	27.30	28.35	27.00
24.30	27.80	27.60	26.30	27.40	23.50	29.60	27.80
27.60	25.35	27.55	29.00	24.10	27.00	24.50	27.25
26.15	29.30	23.10	27.10	28.50	27.45	26.15	28.35
27.95	25.55	27.55	26.60	24.25	30.00	28.55	28.00
27.30	27.90	25.25	24.10	27.45	24.55	26.55	27.55
26.75	31.00	24.00	25.35	26.50	28.30	27.95	25.55
30.25	28.55	26.75	24.60	25.75	26.55	27.80	28.90
29.55	30.00	24.60	25.75	26.30	27.00	28.25	25.25
25.75	26.25	26.30	26.75	27.90	28.30	25.70	26.30
26.60	27.00	30.75	28.60	28.10	23.50	24.75	25.15
26.30	27.25	28.15	29.10	30.10	29.90	28.55	27.30
26.55	27.55	23.00	24.50	22.85	26.55	27.55	28.10
30.70	28.60	27.90	26.80	24.10	25.25	26.30	27.90
26.90	25.30	25.80	28.85	27.55	27.30	25.00	26.00
26.55	27.80	28.60	30.55	29.50	24.10	25.15	27.15
28.10	26.30	27.10	24.60	27.80	26.30	27.90	29.80
24.10	25.15	27.50	24.25	25.70	26.80	30.15	29.30
28.15	28.65	24.55	25.85	26.10	27.00	26.80	27.55
29.00	23.00	28.60	29.30	28.55	28.80	27.55	23.60
26.10	27.15	25.75	26.80	27.15	26.30	28.55	25.80
24.55	25.80	26.75	27.30	27.55	28.25	25.60	26.30
26.85	27.30	28.10	32.00	28.15	26.30	27.75	26.25
28.60	26.00						

FREQUENCY TABLES

While the array presents the figures in a shape much more suitable for study than the haphazard distribution first shown, there is still something to be desired before the mind can readily grasp the full significance of the data. The factory manager may see that the smallest amount earned during the week was $22.85, that the largest amount earned was $32.00, and that most of the employees earned between $25.00 and $29.00, but this is still a vague description of the data. By a process of grouping, that is, by putting into common classes all individuals whose earnings fall within certain limits, a simplified and more compact

ARRAY: WEEKLY EARNINGS OF 210 EMPLOYEES

$22.85	$25.15	$26.15	$26.75	$27.45	$27.95	$28.60
23.00	25.15	26.15	26.75	27.45	27.95	28.65
23.00	25.15	26.25	26.80	27.50	28.00	28.70
23.10	25.25	26.25	26.80	27.55	28.10	28.80
23.40	25.25	26.25	26.80	27.55	28.10	28.85
23.50	25.25	26.30	26.80	27.55	28.10	28.90
23.50	25.25	26.30	26.85	27.55	28.10	29.00
23.60	25.30	26.30	26.90	27.55	28.15	29.00
24.00	25.30	26.30	27.00	27.55	28.15	29.10
24.10	25.35	26.30	27.00	27.55	28.15	29.20
24.10	25.35	26.30	27.00	27.55	28.20	29.30
24.10	25.55	26.30	27.00	27.55	28.25	29.30
24.10	25.55	26.30	27.00	27.60	28.25	29.30
24.10	25.60	26.30	27.10	27.60	28.30	29.50
24.15	25.70	26.30	27.10	27.60	28.30	29.55
24.25	25.70	26.30	27.15	27.70	28.30	29.60
24.25	25.75	26.35	27.15	27.75	28.30	29.75
24.30	25.75	26.40	27.15	27.75	28.35	29.80
24.35	25.75	26.50	27.20	27.80	28.35	29.90
24.50	25.75	26.55	27.25	27.80	28.50	30.00
24.50	25.75	26.55	27.25	27.80	28.55	30.00
24.55	25.75	26.55	27.30	27.80	28.55	30.10
24.55	25.80	26.55	27.30	27.80	28.55	30.15
24.55	25.80	26.55	27.30	27.80	28.55	30.25
24.60	25.80	26.60	27.30	27.80	28.55	30.55
24.60	25.85	26.60	27.30	27.90	28.60	30.60
24.60	26.00	26.60	27.30	27.90	28.60	30.70
24.75	26.00	26.70	27.30	27.90	28.60	30.75
24.75	26.10	26.75	27.40	27.90	28.60	31.00
25.00	26.10	26.75	27.40	27.90	28.60	32.00

presentation of the wage distribution may be obtained. The following table shows the results of this grouping process when the range of each class (the *class-interval*) is two dollars.

This table presents a condensed summary of the original figures, a summary which not only gives us the approximate range of the earnings, but shows, also, how the earnings of the 210 workers are distributed throughout this range. There has been a considerable loss of detail, it will be noted.

TABLE 6

Frequency Distribution of Employees

(Classified on the basis of weekly earnings [class-interval = $2])

Weekly earnings	*Number earning stated amount (frequency)*
$22.00 to $23.99	8
24.00 to 25.99	48
26.00 to 27.99	96
28.00 to 29.99	47
30.00 to 31.99	10
32.00 to 33.99	1
	210

From this table we may learn that there are 48 persons who earned during the given week between $24.00 and $25.99, but we cannot learn how the earnings of the 48 individuals were distributed throughout this range of two dollars. All may have earned exactly $24.00, so far as we may know from the figures shown in the table. This loss of detail is an inevitable accompaniment of the condensation and simplification which the process of classification involves.

If the size of the class-interval be decreased the loss of detail is less pronounced, though the increase in the number of classes means a more cumbersome table and one which presents a more complex picture to the eye. The tables on page 55 present the same data, classified with intervals of one dollar, fifty cents, and twenty-five cents.

The four tables we have thus constructed represent four different degrees of condensation of the same data. Tables 6, 7, and 8 present the same general characteristics: a small number of cases in the extreme classes and a more or less regular increase in the frequencies as the center of each of the distributions is approached. The departure from regularity becomes greater the greater the number of classes. Table 9, in which the class-interval is 25 cents, has 38 classes. In this table the distribution of cases throughout the range is highly irregular, with pronounced departures from sym-

Frequency Distributions of Employees

(Classified on the basis of weekly earnings)

Table 7 (Class-interval = $1)		Table 8 (Class-interval = 50 cents)		Table 9 (Class-interval = 25 cents)	
Weekly earnings	*Frequency*	*Weekly earnings*	*Frequency*	*Weekly earnings*	*Frequency*
$22.00 to $22.99	1	$22.50 to $22.99	1	$22.75 to $22.99	1
23.00 to 23.99	7	23.00 to 23.49	4	23.00 to 23.24	3
24.00 to 24.99	21	23.50 to 23.99	3	23.25 to 23.49	1
25.00 to 25.99	27	24.00 to 24.49	11	23.50 to 23.74	3
26.00 to 26.99	42	24.50 to 24.99	10	23.75 to 23.99	0
27.00 to 27.99	54	25.00 to 25.49	12	24.00 to 24.24	7
28.00 to 28.99	34	25.50 to 25.99	15	24.25 to 24.49	4
29.00 to 29.99	13	26.00 to 26.49	22	24.50 to 24.74	8
30.00 to 30.99	9	26.50 to 26.99	20	24.75 to 24.99	2
31.00 to 31.99	1	27.00 to 27.49	24	25.00 to 25.24	4
32.00 to 32.99	1	27.50 to 27.99	30	25.25 to 25.49	8
	210	28.00 to 28.49	17	25.50 to 25.74	5
		28.50 to 28.99	17	25.75 to 25.99	10
		29.00 to 29.49	7	26.00 to 26.24	6
		29.50 to 29.99	6	26.25 to 26.49	16
		30.00 to 30.49	5	26.50 to 26.74	10
		30.50 to 30.99	4	26.75 to 26.99	10
		31.00 to 31.49	1	27.00 to 27.24	11
		31.50 to 31.99	0	27.25 to 27.49	13
		32.00 to 32.49	1	27.50 to 27.74	14
			210	27.75 to 27.99	16
				28.00 to 28.24	9
				28.25 to 28.49	8
				28.50 to 28.74	14
				28.75 to 28.99	3
				29.00 to 29.24	4
				29.25 to 29.49	3
				29.50 to 29.74	3
				29.75 to 29.99	3
				30.00 to 30.24	4
				30.25 to 30.49	1
				30.50 to 30.74	3
				30.75 to 30.99	1
				31.00 to 31.24	1
				31.25 to 31.49	0
				31.50 to 31.74	0
				31.75 to 31.99	0
				32.00 to 32.24	1
					210

metry. The structure of each of the other tables is orderly and approaches more closely a condition of symmetry. Each presents the wage data in condensed and compact form, so that one consulting the tables may learn of the size and distribution of weekly earnings in the factory in question

much more readily than by reference to the chaotic collection of figures first shown. Such organized collections of data are termed *frequency distributions*, and their purpose, as the term implies, is to show in a condensed form the nature of the distribution of a variable quantity throughout the range covered by the values of the variable. The construction of such a table is the first step to be taken in the organization and analysis of quantitative data of the type represented above.

STEPS IN THE CONSTRUCTION OF A FREQUENCY TABLE

This general introduction to the subject of frequency tables has left untouched many important matters in connection with their construction. It remains to present a summary statement of these details. It will be clear that the first step here taken, the arrangement of the items in order of magnitude, is unnecessary in the actual construction of such a table. Having determined the upper and lower limits through an inspection of the data, one has but to decide on the number of classes desired, write the class-intervals on an appropriate blank sheet, and proceed to tally the cases falling in each of the classes thus set off. When this process is completed the frequencies are computed and the totals arranged in tabular form of the type illustrated above. These simple operations involve decisions on a number of points, however.

SIZE OF CLASS-INTERVAL

In deciding upon the size of the class-interval (which is equivalent to deciding upon the number of classes) one fundamental consideration should be borne in mind, namely, that classes should be so arranged that there will be no material departure from an even distribution of cases within each class. This arrangement is necessary because, in interpreting the frequency table and in subsequent calculations based upon it, the mid-value of each class is taken to repre-

sent the values of all cases falling in that class. Thus, in basing calculations upon Table 8, it is assumed that the 22 cases falling between \$26.00 and \$26.50 may all be represented by the mid-value of that class, \$26.25. This assumption will seldom be strictly valid. In the case just cited reference to the original figures will show that it is not a correct assumption. Absolute accuracy would only be obtained by having a class for every value represented in the original figures. Since condensation is necessary an arrangement of classes should be secured which will minimize the error involved, without transgressing other requirements. Table 6 furnishes an example of class-intervals too wide for the material.

The requirement which has just been described clearly calls for a large number of classes. A second requirement, which ordinarily conflicts with this, is that the number of classes should be so determined that an orderly and regular sequence of frequencies is secured. If the classification is too narrow for the data regularity will not be attained in this respect, and a table without structure or order will be secured. Table 9 fails to meet this requirement, as has been pointed out. It is desirable, also, that the number of classes be limited in order that the data may be easily manipulated and their significance readily grasped.

A useful procedure for approximating a suitable class-interval has been suggested by H. A. Sturges. Given a series of N items of known range, a suitable class-interval i may be approximated from the formula

$$i = \frac{\text{Range}}{1 + 3.322 \log N}.$$

The specific figure secured in a given instance is likely to be a fractional value, quite unsuited to actual use. An appropriate round number close to the theoretical value, may be chosen.[1] Thus, in the example cited above, with a

[1] This formula, and the justification for its use, are discussed in "The Choice

range of \$9.15 and N equal to 210, the use of a class-interval of \$1.05 is indicated by the formula. The nearest round number, suitable with reference to other considerations as well, is \$1.00. Table 7, in which this class-interval is employed, seems to conform most thoroughly to all the requirements we have set forth.

LOCATION OF CLASS LIMITS

The location of class limits is a matter of considerable importance, for attention to this matter will simplify tabulation and facilitate later calculation. Tabulation of data is easiest when class limits are integers and the class-interval itself is a whole number. Calculation of averages and other statistical measures is facilitated when the mid-values of classes are integers. Suitable class limits and mid-points are usually secured when the data permit class-intervals of 5 or multiples of 5 to be employed, though such an arrangement is by no means essential.

Some types of data show a tendency to cluster or concentrate about certain values on the scale along which they are distributed. This is illustrated by the following figures which form part of a table showing the number of pieces of commercial paper discounted by the Federal Reserve Banks in 1921, distributed according to rates of discount or interest charged by member banks:

Rate (per cent)	*Number of pieces*
6	18,970
6¼	697
6½	4,616
6¾	135
7	17,362
7¼	10

of a Class Interval" by Herbert A. Sturges, *Journal of the American Statistical Association*, March, 1926, 65–6. The use of the formula rests on the assumption that the proper distribution into classes is given, for all numbers which are powers of 2, by a series of binomial coefficients. The relation of the terms in the binomial expansion to the theory of frequency distributions is discussed below, in Chapter XIII.

Here is a quite obvious bunching about the integers, with a secondary concentration at each half of one per cent. No cases at all fall between the quarter values here shown. It is clear that in classifying such data the mid-points of the various classes should fall at those values about which the cases are concentrated, and class limits must be located with this end in view. For, as noted above, calculations based upon the frequency table are performed upon the assumption that all the items in each class are concentrated at the mid-point of that class. Thus, if a class interval of one half of one per cent were selected in the above example, the classes should extend from $5\frac{3}{4}$ to (but not including) $6\frac{1}{4}$, $6\frac{1}{4}$ to $6\frac{3}{4}$, etc., rather than from 6 to $6\frac{1}{2}$, $6\frac{1}{2}$ to 7, etc.

ACCURACY OF OBSERVATIONS AND THE DEFINITION OF CLASSES

In the construction of frequency tables it is essential that there be a clear definition of classes, so that there may be no uncertainty as to their range and no question as to the precise class in which a given case falls. A table with an arrangement similar to the following is sometimes encountered:

Class-interval	*Frequency*
0 to 10	3
10 to 20	8
20 to 30	15
30 to 40	6
40 to 50	2

In the absence of explanation, a question arises at once as to whether a case with a value of 10 would fall in the first or in the second class. It is highly desirable that the range of each class be indicated in some such way as the following, in order that this ambiguity may be avoided:

Class-interval	*Frequency*
0 to 9.9	3
10 to 19.9	8
20 to 29.9	15
30 to 39.9	6
40 to 49.9	2

This procedure solves the difficulty, however, only in case the observations are accurate to the nearest tenth. If the observations are accurate only to the nearest unit (that is, if the cases recorded as having a value of 10 actually lie between 9.5 and 10.5) a mere change in the description of the class range does not solve the problem of allocating a case at the class limit. In such a case an observation falling at a class boundary may be cut in two, one half being allocated to each of the adjacent classes.

Yule lays down the useful principle that in fixing a class boundary the limit should be carried to a farther place in decimals, or a smaller fraction, than the values of the individual cases as originally recorded. Thus, in the preceding example, if observations were correct to the nearest tenth, it would mean that a value recorded as 9.9 actually lay between 9.85 and 9.95. In accurately describing the classes, therefore, the intervals should be given as 0 to 9.95, 9.95 to 19.95, etc. (Since the observations to be tabulated are recorded only to the first decimal place no ambiguity arises from the apparent over-lapping of these class limits.) It should be noted that the values of the mid-points, with these class limits, would be 4.95, 14.95, etc. In presenting and using the table as given above the real meaning of the class limits should be borne in mind. In all cases class boundaries must be fixed with reference to the accuracy of the observations.

The work of tabulation is simplified if, in designating a class, both limits are stated, as above. Errors are likely if only the lower limit of each class is given, or if the mid-point alone is designated. It is desirable, however, par-

ticularly if calculations are to be based upon the table, to include a separate column showing the values of the mid-points of the various classes.

OTHER REQUIREMENTS

Class-intervals should be uniform throughout the table in order that all classes may be comparable. Occasionally tables are published with varying class-intervals, so that on one section of the scale the number of items falling within a class having an interval of 5 is given, and on another section of the scale the number of items falling within a class having a range of 10 is given. Obviously, comparison of classes is impossible. It may be desirable to show in more detail the cases falling within certain ranges on the scale, but this end is best achieved by the construction of a supplementary table relating only to the cases falling within this restricted section. The utility of the main table is not lessened thereby.

Similar in nature is the requirement that there should be no indeterminate classes, that is, classes the ranges of which are not defined. Had all the individuals making $30.00 and over in the illustration of piece-work earnings been entered in a class with the designation "$30.00 and over," the upper limit of this class would have been quite uncertain. This fault in a table is a vital one when it is desired to base calculations upon the data contained in the table. When there are several extreme cases the inclusion of such classes is sometimes unavoidable, but when this is done the actual values of the cases included in such "open end" classes should be given in a footnote to the table.

The errors described in the two preceding paragraphs are exemplified in the table on page 62.

In this case the ranges of the two "open end" classes are not known. The ranges of the intermediate classes vary, being $5.00 for two classes, $10.00 for one class, $20.00 for one class and $25.00 for two classes. The purposes of a

TABLE 10

Frequency Distribution of Rented Dwellings in Reno, Nevada, 1934
(Classified on the basis of rental value[1])

Monthly rental	*Number of dwellings in each class (frequency)*
Under $10.00	327
$10.00 to $14.99	349
15.00 to 19.99	521
20.00 to 29.99	1,039
30.00 to 49.99	1,075
50.00 to 74.99	189
75.00 to 99.99	24
$100.00 and over	9
	3,533

special investigation may sometimes be served by the use of such a form, but a table of this type is poorly adapted to the requirements of statistical calculation.

THE STRUCTURE OF STATISTICAL TABLES

The preceding discussion has been confined to certain more or less technical problems which arise in the construction of a frequency table. Nothing has been said directly as to the form of the completed table, the arrangement of columns and rows, the title, the notation. No general principles of tabular arrangement have been laid down. While no detailed treatment of these principles is possible within the scope of the present discussion, certain general considerations relating to the structure of statistical tables may be suggested.

The statistical table is merely a device for presenting in summary fashion a mass of quantitative data. Unless the summary be clear, significant, concise, and readily interpreted nothing has been gained by the process of tabulation

[1] The table is taken from *Real Property Inventory, 1934. Summary and Sixty-Four Cities Combined,* Department of Commerce, Washington. Figures for 255 rented dwellings in Reno were not reported.

and classification. A sprawling, formless table is like a rambling, unintelligible discourse. There must be a purpose in back of each table, and this purpose should be clearly brought out in its arrangement. The means by which this purpose may be attained in a given case must be determined with reference to the specific conditions affecting that case, but standard practices should be followed, in so far as possible. The following general principles will be found helpful in deciding upon the form and arrangement of statistical tables:

1. The title should constitute a clear, concise and complete description of the material assembled in the table.
2. Headings of columns and rows should be concise and unambiguous.
3. Variable quantities should increase from left to right and from top to bottom, when such arrangement is feasible.
4. Columns and rows may be numbered to facilitate reference to the table.
5. The units of measurement employed should be clearly indicated.
6. Sources should be given in all cases.
7. The table should constitute a unit, self-sufficient and self-explanatory. All explanations necessary for its interpretation should be included as integral parts of the table, or in the form of footnotes.

Graphic Representation of Frequency Distributions

Frequency distributions of the type illustrated above serve a very important statistical function in presenting a compact summary of data, and in preparing these data for further manipulation. Such distributions may be presented not only in tabular form, but graphically, utilizing the general principles of the coördinate system which were explained above. Many of the characteristic features of a frequency distribution are most clearly revealed when the graphic method is adopted.

Table 6, presenting the weekly earnings of 210 employees, with a class-interval of two dollars, is depicted graphically

in Fig. 23. In this figure class-intervals are plotted along the x-axis and the corresponding class-frequencies along the y-axis, appropriate scales being selected. The fact should be noted that the scale of abscissas starts not with zero, but with \$20. For convenience in presentation that part

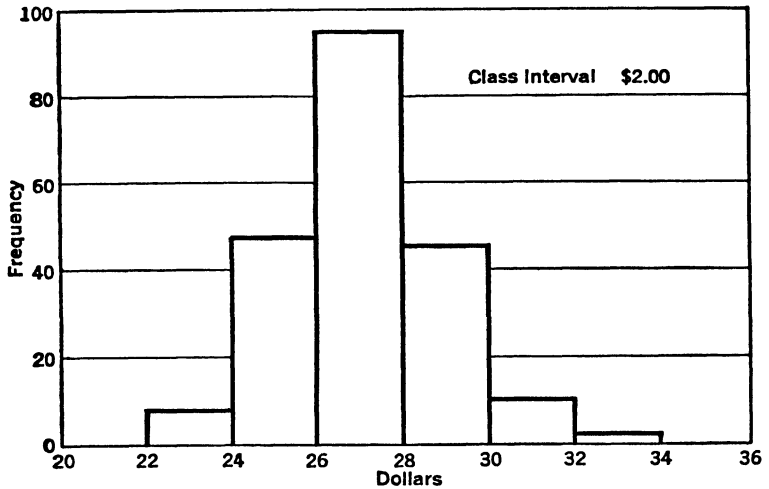


FIG. 23. — Column Diagram: Distribution of 210 Employees Classified on the Basis of Weekly Earnings (Class-interval = \$2.00)

of the scale extending from 0 to \$20 is omitted. The student should bear this in mind in seeking to secure a correct impression of the relations between the two variables plotted. In constructing such a figure, which is termed a *column diagram* or *histogram*, short horizontal lines are drawn connecting the points plotted to represent the upper and lower limits of each class-interval. In interpreting this diagram it should be noted that the areas of the different rectangles are proportional to the number of cases represented, the total area representing the entire 210 cases. This device thus presents to the eye a very clear picture of the distribution, showing quite unmistakably the relative number of workers falling in each of the wage classes.

The classes in this case are so large, however, that some

violence is done to the facts. So many details are lost that a true conception of the disposition of the items is not given. Fig. 24 is a histogram depicting the distribution of cases when a class-interval of one dollar is used. In this case, with smaller steps, we approach more closely an orderly and symmetrical distribution. The same is true of Fig. 25

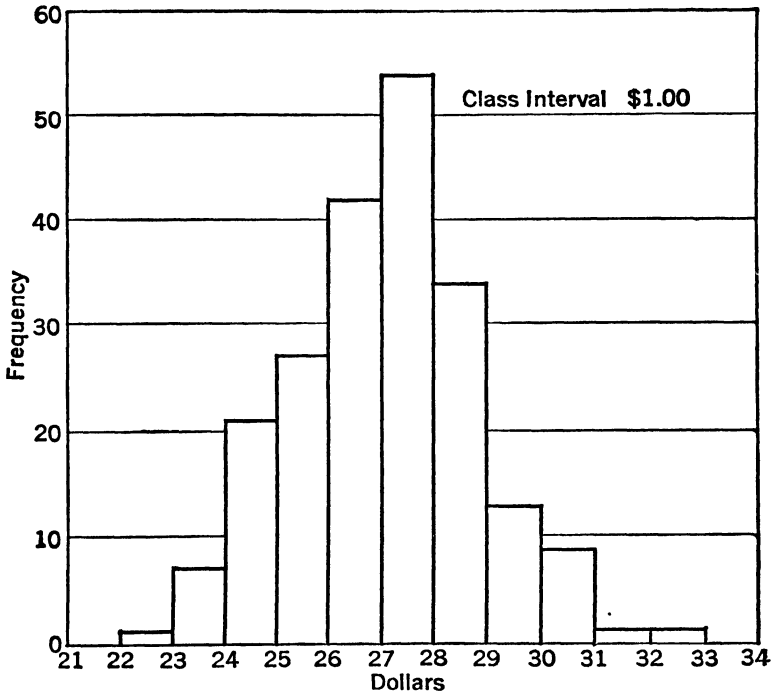


FIG. 24. — Column Diagram: Distribution of 210 Employees Classified on the Basis of Weekly Earnings (Class-interval = $1.00)

which shows the distribution when the class-interval is fifty cents. The distribution represented in Fig. 26 has a class-interval of twenty-five cents which, as has been pointed out, is too narrow for the data, with the result that a quite irregular structure is secured. (It should be noted that the vertical scale is not the same in these four figures, so that comparison with respect to class-frequencies is only possible by reference to the scale figures.)

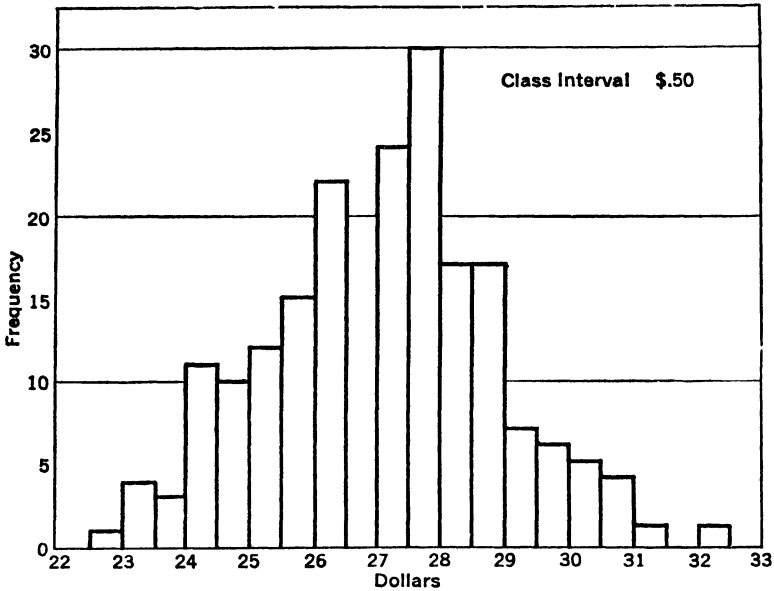


FIG. 25. — Column Diagram: Distribution of 210 Employees Classified on the Basis of Weekly Earnings (Class-interval = $.50)

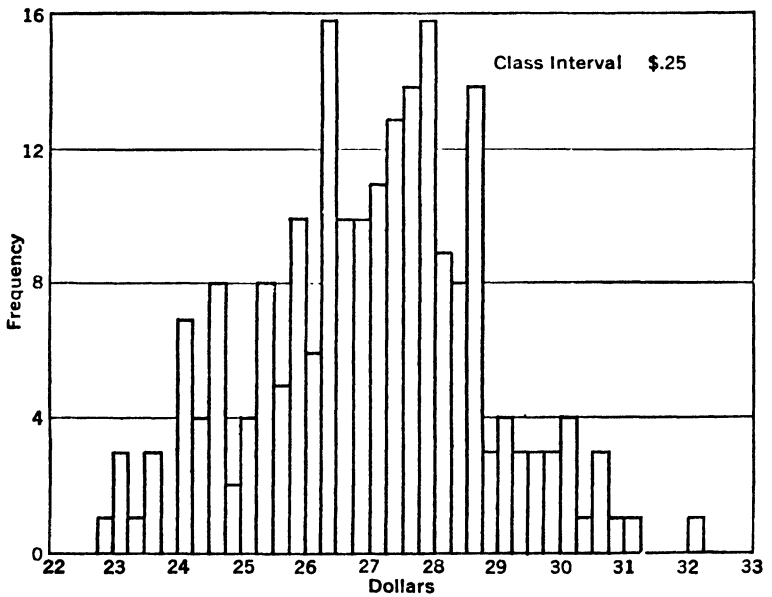


FIG. 26. — Column Diagram: Distribution of 210 Employees Classified on the Basis of Weekly Earnings (Class-interval = $.25)

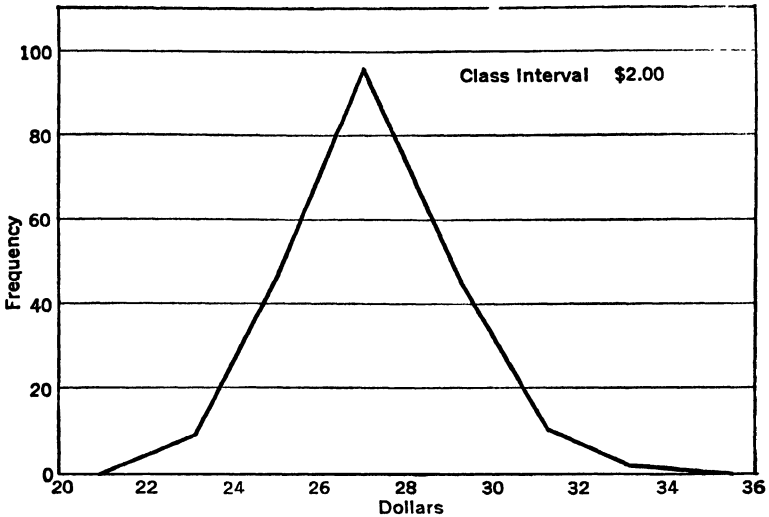


FIG. 27. — Frequency Polygon: Distribution of 210 Employees Classified on the Basis of Weekly Earnings (Class-interval = $2.00)

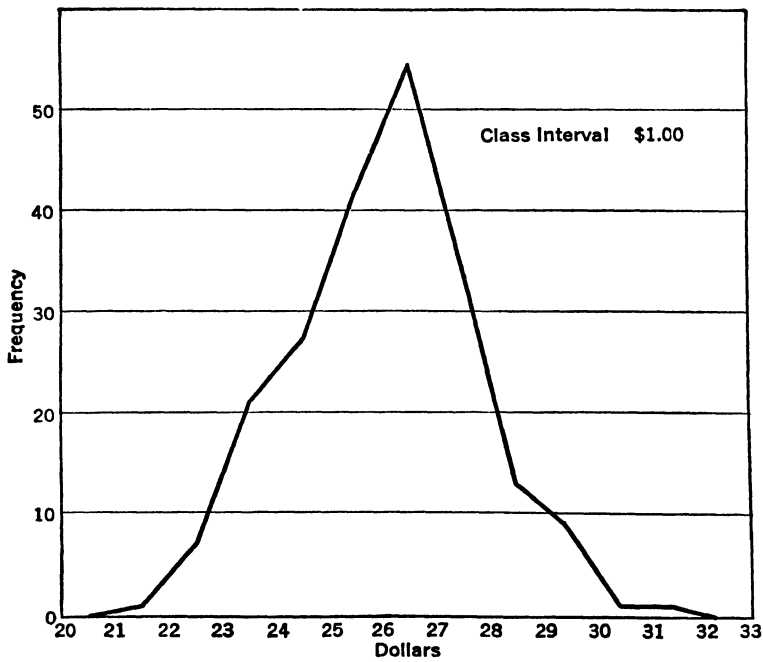


FIG. 28. — Frequency Polygon: Distribution of 210 Employees Classified on the Basis of Weekly Earnings (Class-interval = $1.00)

Frequency polygons corresponding to the histograms of Figs. 23, 24, and 25 are shown in Figs. 27, 28, and 29. Each of these polygons has been constructed by plotting as abscissas the mid-points of the class-intervals, and as ordinates the class-frequencies, the points thus secured being

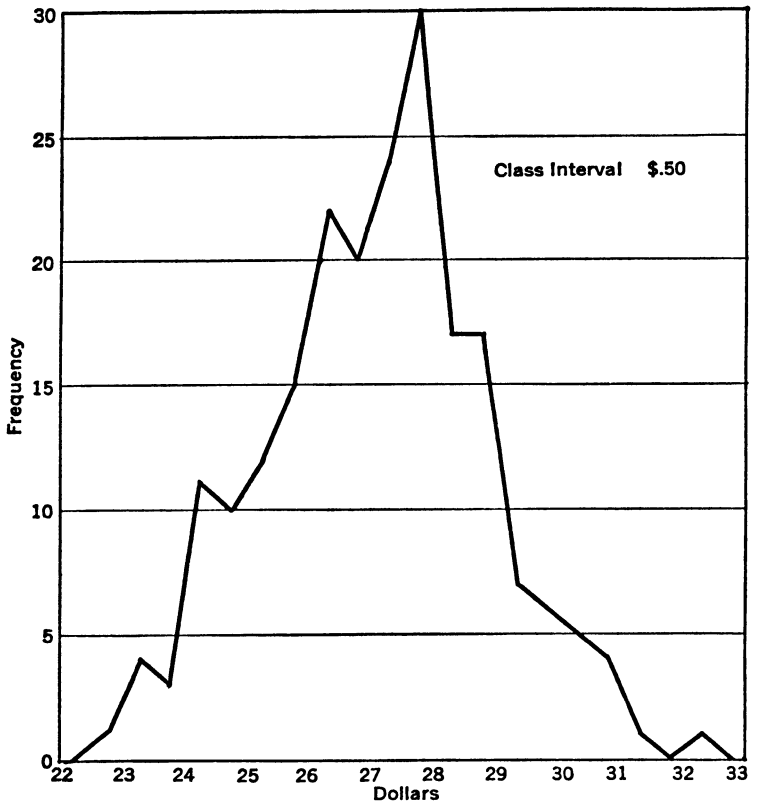


FIG. 29. — Frequency Polygon: Distribution of 210 Employees Classified on the Basis of Weekly Earnings (Class-interval = $.50)

connected by a broken line. In completing such a figure the class next below the lowest one on the scale and the class next above the highest one on the scale are included, the class-frequency being zero in each case. The ends of the polygon thus connect with the base line at the mid-

points of these two extra classes. In the case of the frequency polygon the entire area under the curve represents the entire number of cases, but the area of a given interval cannot be taken to be proportional to the number of cases in that interval, because of irregularities in the distribution on either side of the given class. The heights of the ordinates at the mid-points of the various classes are, of course, scaled to represent the class-frequencies.

The Smoothing of Curves

Attention is again called to the results secured with varying class-intervals. As the class-interval is decreased, up to a certain point, the histograms and polygons become smoother and more regular. Beyond that point breaks begin to appear in the data; the regular change in class-frequencies which was found when the classes were larger is broken by the appearance of irregular classes which seem to depart from the general rule. In Fig. 25 these have become quite pronounced. Such irregularities, it is obvious, are exceptions to a general rule which seems to prevail, the general rule that the numbers of workers falling within the different wage classes increase from the lower limit of earnings up to a maximum in the neighborhood of $27.50, and then decrease till, at the upper limit of $32, but one worker is found. Since all the 210 individuals are engaged in the same work, and since their earnings depend only upon their rapidity and skill, one would expect a quite regular increase and decrease. If we had figures not for one week only, but for 52 weeks, and took the average weekly earnings of each of the 210 workers for the year, we should expect greater regularity with the smaller class-intervals than is actually found, since the accidental fluctuations peculiar to one week alone would thus be eliminated. Or, if we had earnings during one week for 10,920 workers (52 times 210), the same result would be secured. Thus, if regularity and smoothness are to be secured, it is

essential not only to decrease the size of the classes but also to increase the number of cases, in order that the accidental irregularities which affect a small number of observations may be eliminated. A refined classification with a small number of cases leads to the condition exemplified in Fig. 26. But such an increase in the number of cases is, in general, a practical impossibility. We wish, if possible, to develop a feasible method of approximating the distribution which would be secured with very small class-intervals and a very large number of cases. Such an approximation is possible through the device of curve-smoothing. By this method we may secure a smooth *frequency curve* which lacks the irregularities occasioned by minor fluctuations.

Such a smooth frequency curve serves to represent the true underlying distribution of the data. It was pointed out that areas in the frequency polygon are not proportional to the number of cases included, the cause lying in the irregularities of the data. In a smoothed frequency curve these irregularities have been eliminated, and the area between ordinates erected at given points on the scale of abscissas is assumed to be proportional to the theoretical frequency of cases between the given values. Moreover, a smooth trend having been established, frequencies for intermediate values not shown in the original table may be determined by interpolation.[1]

The following data,[2] representing the distribution in 1918 of personal incomes below $4,000, will serve to exemplify the smoothing process.

[1] The limitations of practical statistical work are such that there must of necessity be many gaps in the data. The given values of the variables are not continuous. Interpolation is the process of estimating values of a variable quantity between given values, or of locating a point on a curve between given points. That interpolation is most accurate which leads to estimated values having the highest degree of consistency with the given values.

[2] From Vol. I, *Income in the United States*, National Bureau of Economic Research. New York, Harcourt, Brace & Co., 1921, 132–33.

TABLE 11

Distribution of Income among Personal Income Recipients in 1918

(Including all personal incomes below $4,000)

Income class[1]	*Number of persons*[2]
$ 0 to $ 100	62,809
100 to 200	103,704
200 to 300	209,087
300 to 400	489,963
400 to 500	961,991
500 to 600	1,549,974
600 to 700	2,154,474
700 to 800	2,668,466
800 to 900	3,013,034
900 to 1,000	3,144,722
1,000 to 1,100	3,074,351
1,100 to 1,200	2,850,526
1,200 to 1,300	2,535,285
1,300 to 1,400	2,205,728
1,400 to 1,500	1,832,230
1,500 to 1,600	1,512,649
1,600 to 1,700	1,234,397
1,700 to 1,800	999,996
1,800 to 1,900	811,236
1,900 to 2,000	663,789
2,000 to 2,100	549,787
2,100 to 2,200	463,222
2,200 to 2,300	395,115
2,300 to 2,400	340,141
2,400 to 2,500	295,490
2,500 to 2,600	258,650
2,600 to 2,700	227,731
2,700 to 2,800	201,488
2,800 to 2,900	178,901
2,900 to 3,000	154,499
3,000 to 3,100	142,802
3,100 to 3,200	128,217
3,200 to 3,300	115,583
3,300 to 3,400	104,504

[1] The definition of classes used is equivalent to "$0 to and not including $100," etc. Thus an individual with an income of $100 would fall in the second class.

[2] The National Bureau's report states "The numbers below are given to the nearest unit. It is not pretended that such arithmetic accuracy is anything more than technical."

TABLE 11—*Continued*

Distribution of Income among Personal Income Recipients in 1918

Income class	*Number of persons*
$3,400 to 3,500	94,803
3,500 to 3,600	86,405
3,600 to 3,700	79,023
3,700 to 3,800	72,562
3,800 to 3,900	66,900
3,900 to 4,000	61,894

Figures 30, 31, and 32 present column diagrams of these income data, grouped with class-intervals of $500, $200, and

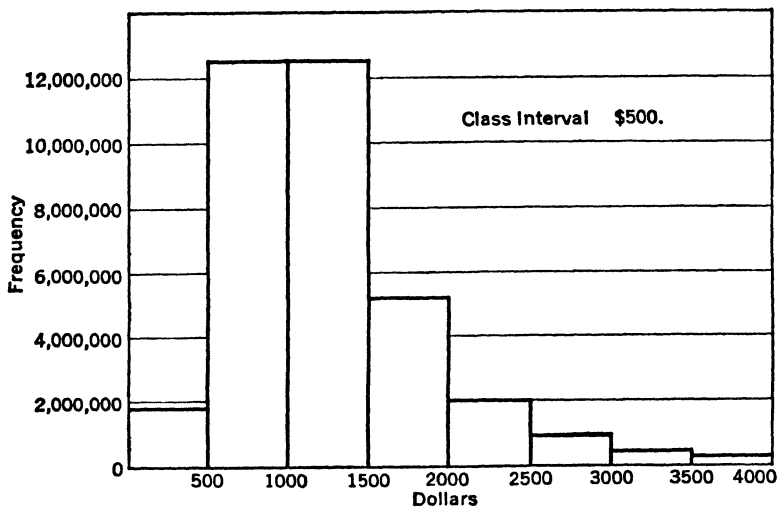


FIG. 30. — Column Diagram: Distribution of Personal Income Recipients in the United States, 1918. Including all Recipients of Incomes below $4,000 (Class-interval = $500)

$100. As the class-interval is decreased the histograms become more regular and uniform, but our original data permit us to carry this process only to the point where the class-interval is $100. Our problem is to determine the underlying distribution which the data approximate more and more closely as the class-interval is lessened. If we replace the broken line of the histogram by a smooth curve

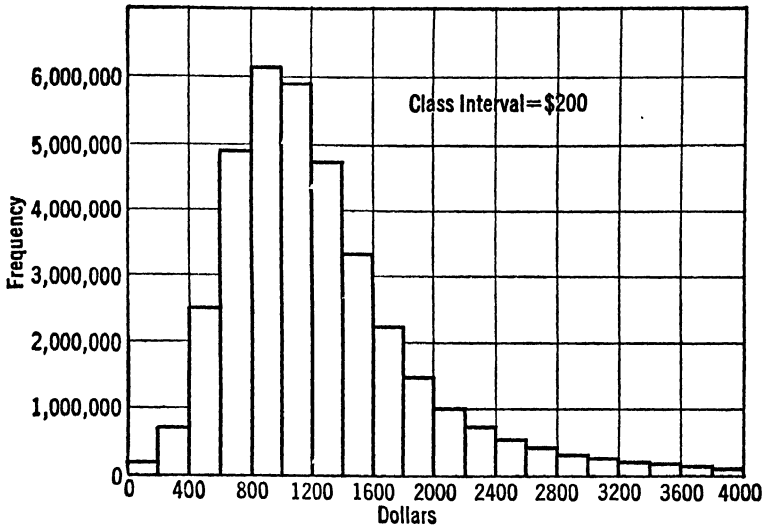


FIG. 31. — Column Diagram: Distribution of Personal Income Recipients in the United States, 1918. Including all Recipients of Incomes below $4,000 (Class-interval = $200)

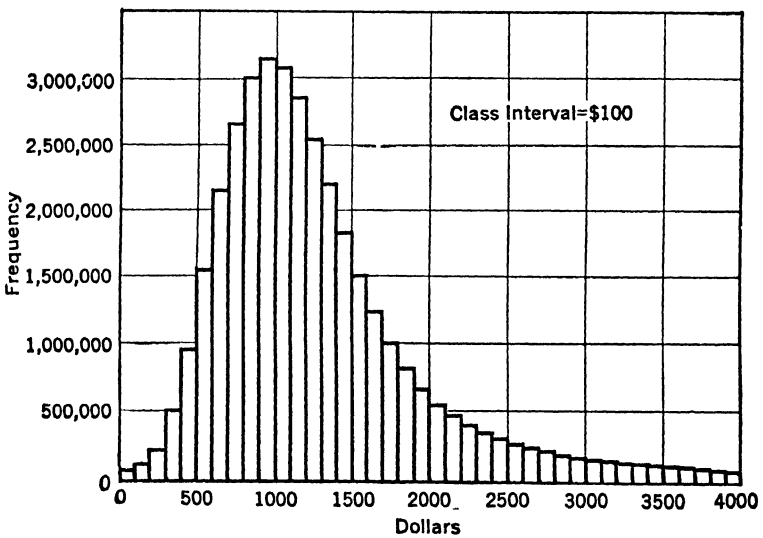


FIG. 32. — Column Diagram: Distribution of Personal Income Recipients in the United States, 1918. Including all Recipients of Incomes below $4,000 (Class-interval = $100)

enclosing the same total area as the histogram and so drawn through the points of the histogram that *the area cut from each rectangle is approximately equal to the area added to the same rectangle by the curve,* we will have a frequency

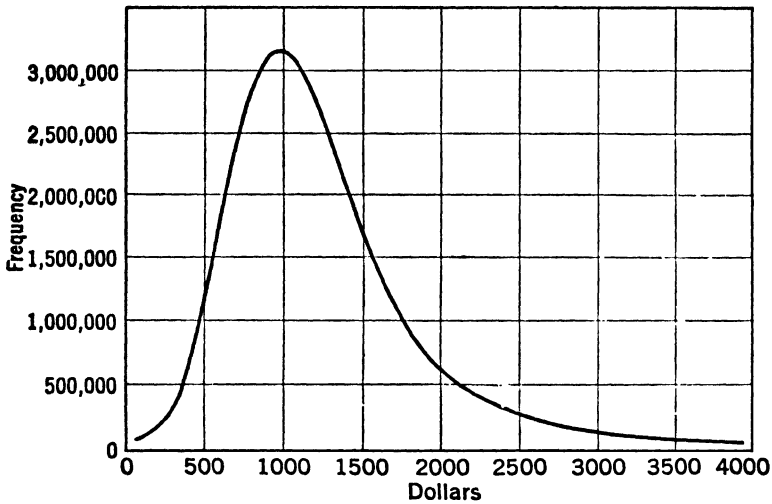


FIG. 33. — Frequency Curve: Distribution of Personal Income Recipients in the United States, 1918. Including all Recipients of Incomes below $4,000 (Derived from the column diagram with class-interval of $100)

curve representing the desired distribution. The requirement that the same total area be enclosed is fundamental. Exceptions to the rule concerning the area of individual rectangles will frequently occur because of the existence of quite irregular classes, but as a general working principle it is helpful. (More refined methods of fitting a smooth curve to data will be discussed at a later point, but a process of smoothing by inspection such as that described above gives a fairly close approximation to the required curve.)

Figure 33 illustrates the result of smoothing the histogram of income distribution shown in Fig. 32. Here the quite artificial jumps between income classes are smoothed out, and we secure the graduation by infinitesimal increments which we should expect to find when the incomes of

so many millions of persons are included. Here we have that which we desired — an approximation to the true underlying distribution, with the sharp breaks resulting from the method of classification eliminated.

CONTINUOUS AND DISCRETE SERIES

The logical validity of the smoothing process is dependent upon the nature of the data being manipulated. From this point of view frequency series of the type discussed above may be divided into two classes, *continuous series* and *non-continuous* or *discrete series*. A continuous series is one in which the values of the independent variable increase or decrease by increments which are infinitely small. A discrete series is one in which the phenomena represented by the independent variable always change in value by definite amounts. The curve of underlying values rises not smoothly, as for the continuous series, but by jumps.

The fact should be emphasized that in making this distinction we are speaking of the values as they would be found in the underlying universe of phenomena from which the actual bodies of material we study are drawn. Any given sample, whether representing continuous or discrete series, will be marked by breaks in the values of the independent variable. This will be true, in the case of a continuous series, because of the limitations upon the instruments and senses we use in measuring. Thus if the heights of 100 men be measured, the independent variable of the frequency series (height) will increase by finite amounts. We may measure to the nearest inch, or perhaps to the nearest eighth of an inch. Yet if ten thousand or ten million men were arranged in order of height the differences between successive individuals would be infinitely small. Height is a continuous variable, even though the values found in a given sample are marked by discontinuity.

Quite different is the distribution of such a variable as interest or discount rates. If one were to secure 100 such

quotations and rank them in the order of size the variations would be discontinuous, as in the sample of men whose heights were measured. But in the case of heights the underlying values, if they could be determined for a large population, would be marked by continuous variation, whereas, were an infinite number of discount rate quotations secured, there would still be breaks in the sequence. Discount rates increase or decrease by one quarter or one half of one per cent, not by infinitesimal amounts. Such a series is termed discrete, or non-continuous.

The smoothing process provides a means of securing an approximation to the distribution of values as they would be found if a sample could be increased indefinitely in size. It is based upon the assumption that the irregularities found in the sample actually studied are accidental, and that the underlying values would show continuous and unbroken variation. Obviously, therefore, it is only fully justified when applied to a continuous series. A histogram of human heights may be smoothed in order to secure a representation of the true underlying distribution in the population at large, and interpolation based upon this smoothing process is valid. But smoothing is quite illogical for a markedly discontinuous series. It would be meaningless to construct a smooth curve showing the distribution of discount rates for the purpose of securing the theoretical frequency of a rate of 4.3675 per cent. In practical statistical work, however, it is frequently helpful to handle discrete series as though they were continuous, and in these cases the smoothing device may be employed. But in the interpretation and use of the smoothed curve the important logical distinction between continuous and discontinuous variation should be kept clearly in mind.

Cumulative Arrangement of Statistical Data

For certain purposes it is desirable to arrange data cumulatively, rather than in separate and exclusive classes of

the type illustrated in the frequency tables presented above. The following material will illustrate some of the advantages of this arrangement.

In a study of the durability of telephone poles [1] these results were secured:

TABLE 12

Frequency Distribution of 248,707 Telephone Poles, Classified According to Length of Life

Length of life (years)	*Number of poles (frequency)*
0– 0.9	1,150
1– 1.9	4,221
2– 2.9	10,692
3– 3.9	13,966
4– 4.9	16,633
5– 5.9	18,211
6– 6.9	19,011
7– 7.9	19,260
8– 8.9	20,909
9– 9.9	19,879
10–10.9	20,764
11–11.9	15,454
12–12.9	14,237
13–13.9	13,779
14–14.9	9,764
15–15.9	8,534
16–16.9	7,659
17–17.9	6,918
18–18.9	4,591
19–19.9	1,798
20–20.9	815
21–21.9	313
22–22.9	102
23–23.9	47

The table shows that 1,150 poles were scrapped during the first year of use, that 4,221 were scrapped after reaching the age of one year and before reaching the age of two

[1] "Replacement Insurance," Edwin Kurtz. *Administration*, July, 1921, 41–69.

years, and so on. This is simply a frequency table of the ordinary type. A much more significant arrangement for many purposes is secured when the figures are assembled cumulatively, as in the following table.

TABLE 13

Cumulative Distribution of 248,707 Telephone Poles, Classified According to Length of Life

(Cumulated upward with reference to life scale)

Length of life	*Number of poles surviving (frequency)*
Less than 1 year	1,150
" " 2 years	5,371
" " 3 "	16,063
" " 4 "	30,029
" " 5 "	46,662
" " 6 "	64,873
" " 7 "	83,884
" " 8 "	103,144
" " 9 "	124,053
" " 10 "	143,932
" " 11 "	164,696
" " 12 "	180,150
" " 13 "	194,387
" " 14 "	208,166
" " 15 "	217,930
" " 16 "	226,464
" " 17 "	234,123
" " 18 "	241,041
" " 19 "	245,632
" " 20 "	247,430
" " 21 "	248,245
" " 22 "	248,558
" " 23 "	248,660
" " 24 "	248,707

It is important to note that it is possible to cumulate a frequency series in two different ways. From the above table we may determine readily the number failing to attain any given age. It is often more convenient to reverse the process, so that the table will enable the total number above any given value to be immediately determined. When

the telephone pole figures are thus *cumulated downward* the following table is secured.

TABLE 14

Cumulative Distribution of 248,707 *Telephone Poles, Classified According to Length of Life*

(Cumulated downward with reference to life scale)

(1) *Length of life*	(2) *Number of poles surviving (frequency)*	(3) *Per cent*
0 and more	248,707	100.0
1 year " "	247,557	99.5
2 years " "	243,336	97.8
3 " " "	232,644	93.6
4 " " "	218,678	88.0
5 " " "	202,045	81.2
6 " " "	183,834	73.8
7 " " "	164,823	66.3
8 " " "	145,563	58.5
9 " " "	124,654	50.1
10 " " "	104,775	42.1
11 " " "	84,011	33.8
12 " " "	68,557	27.6
13 " " "	54,320	21.8
14 " " "	40,541	16.3
15 " " "	30,777	12.4
16 " " "	22,243	8.9
17 " " "	14,584	5.9
18 " " "	7,666	3.1
19 " " "	3,075	1.2
20 " " "	1,277	0.5
21 " " "	462	0.2
22 " " "	149	0.06
23 " " "	47	0.02
24 " " "	0	0.00

Cumulative tables such as those given above have distinct advantages in the handling of many types of data. Life tables are generally presented in this form. The scientific study of depreciation will lead to the construction of elaborate "mortality tables" for various types of equipment, and these will be most useful in the cumulative form. It is frequently desirable to reduce the frequencies to per-

centages, as in column (3) of Table 14, though it should not be forgotten that the significance of the percentages depends upon the absolute numbers upon which they are based.

THE OGIVE, OR CUMULATIVE FREQUENCY CURVE

The general utility of such cumulated data is limited by the classification system necessarily adopted in condensing

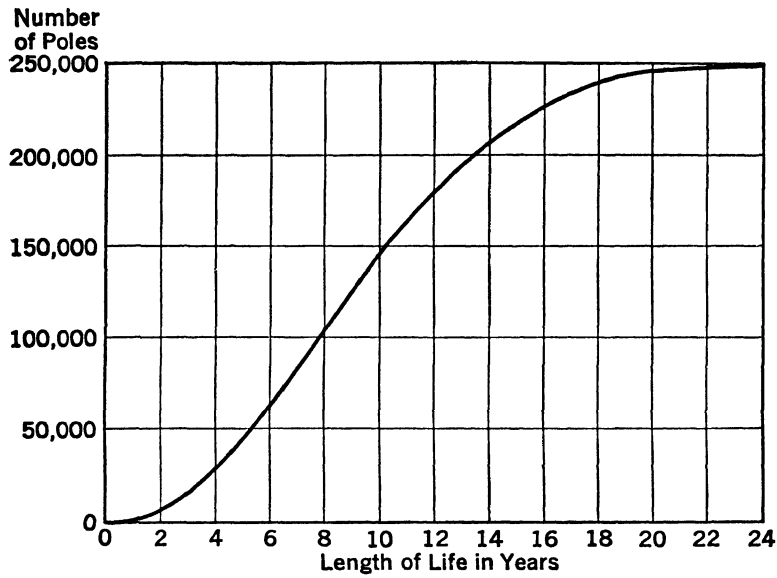


FIG. 34. — Cumulative Frequency Curve: Distribution of Telephone Poles Classified according to Length of Life (Cumulated upward)

the material. Unless we interpolate mathematically we are limited to the points on the scale actually noted in the two tables. For this reason, a generalized cumulative curve similar to the smoothed frequency curve described in the preceding section is desirable. If the values given in Table 13 be plotted on coördinate paper (the length of life in each case as abscissa, and the corresponding number of poles as ordinate) and a smooth curve drawn through the points thus plotted, the cumulative frequency curve shown in

Fig. 34 is secured. In Fig. 35 the data of Table 14 are plotted.

Such a curve constitutes one of the most effective and useful representations of a frequency series. It is obvious that the limitations of the particular class-interval adopted are in large part removed; the shape of the curve will be fundamentally the same, though the class-interval and num-

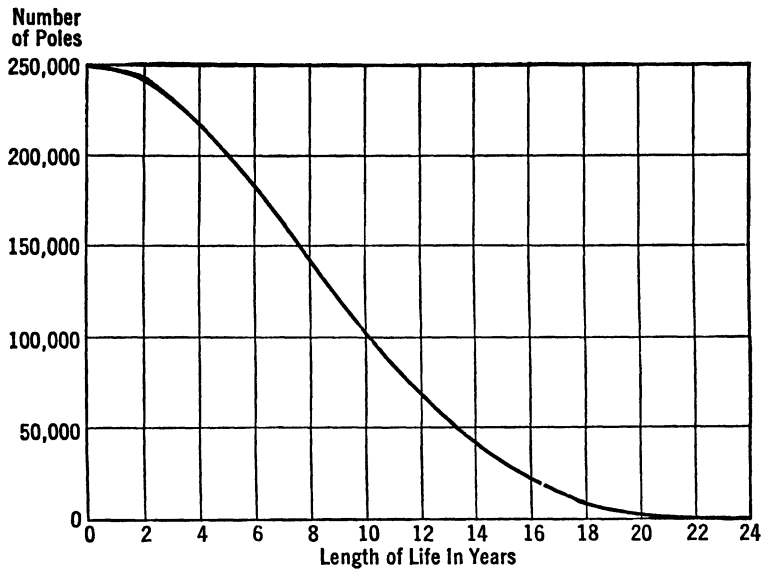


FIG. 35. — Cumulative Frequency Curve: Distribution of Telephone Poles Classified according to Length of Life (Cumulated downward)

ber of classes may vary. Frequency curves of the usual type may not be compared unless the groupings are the same, but cumulative frequency curves are subject to no such restriction. Moreover, uneven class-intervals do not distort the *ogive*, or cumulative curve, as they do the ordinary frequency curve.

The cumulative curve is particularly well adapted to interpolation. Thus if it is desired to know the number of poles surviving less than $15\frac{1}{2}$ years, the value of the ordinate of the curve having $15\frac{1}{2}$ as abscissa may be approxi-

mated from Fig. 34. A value of 222,000 is secured. If the number surviving 8½ years or more is desired, a similar estimate may be made from Fig. 35. The interpolated figure in this case is 135,000.

Another type of interpolation possible with such a curve is the determination of the number of cases falling within any given interval. One is not limited to the class-intervals marked out in the original tables. For instance, it may be desirable to know the number of poles surviving more than 10½ but less than 15 years. Reading from the table or from the chart we find that 217,930 poles survived less than 15 years. Interpolating on the chart in the manner described above a figure of 154,000 is secured for the number surviving less than 10½ years. Subtracting the latter figure from the former we have 63,930 as the number of poles falling within the 10½ to 15 years interval. The figure is, of course, an approximation to the true value, as are all values secured through such smoothing and interpolation.

It should be noted that the ogive may be derived directly from the array, without the formation of a frequency table as an intermediate step. This curve, in fact, may be looked upon as merely a graphic representation of the array. It represents one of the simplest forms of statistical organization, as well as one of the most effective methods of manipulating quantitative data.

RELATION BETWEEN THE OGIVE AND THE FREQUENCY CURVE

The ogive and the frequency curve are merely two different arrangements of precisely the same material, each arrangement having certain distinctive advantages. The characteristics of each may be more clearly apparent if the structural relationship between these two curves is understood. This relationship is graphically portrayed in Fig. 36.[1]

[1] The suggestive arrangement shown in this figure was originated by Robert E. Chaddock.

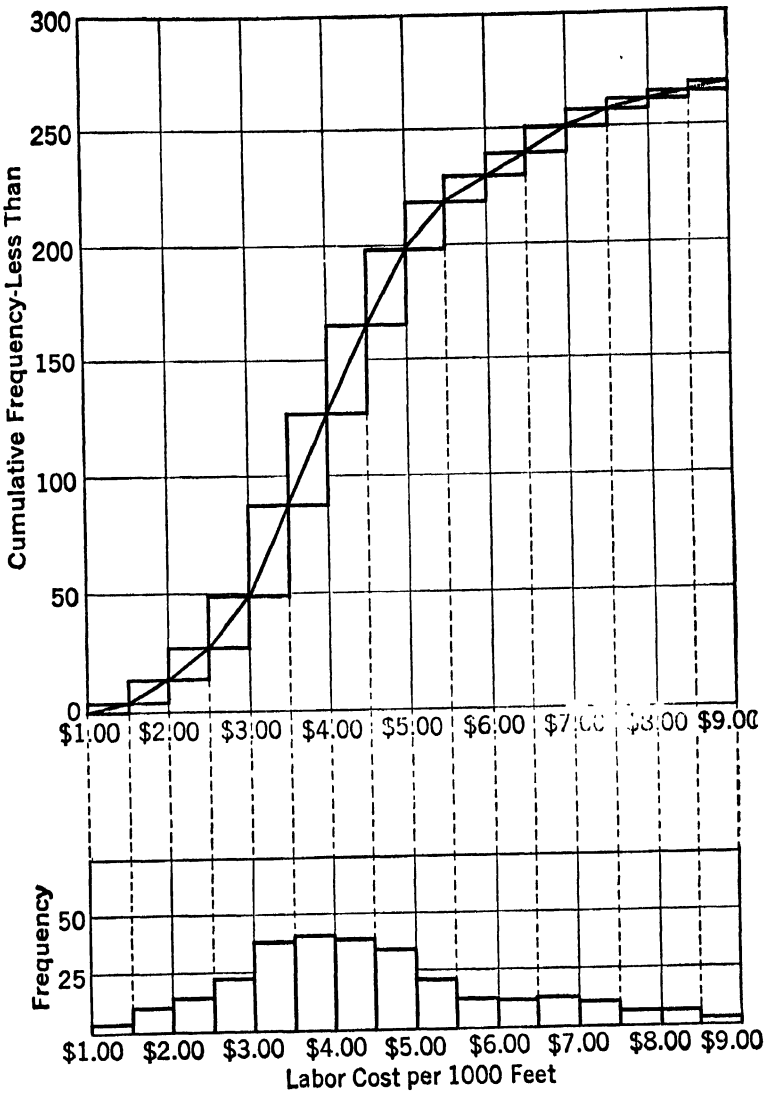


FIG. 36. — Distribution of Sawmills in the United States Classified according to Labor Cost in 1921. Illustrating the Structural Relation between the Ogive and the Frequency Curve.

This figure is based upon the following frequency table, showing the distribution of sawmills in the United States, classified on the basis of labor cost per 1,000 feet of lumber produced.[1]

TABLE 15

Frequency Distribution of 269 *Sawmills in the United States Classified According to Labor Cost in* 1921

Labor cost (all employees) per 1,000 feet, board measure	*Number of establishments (frequency)*
$1.00–$1.49	3
1.50– 1.99	10
2.00– 2.49	14
2.50– 2.99	22
3.00– 3.49	38
3.50– 3.99	40
4.00– 4.49	38
4.50– 4.99	33
5.00– 5.49	20
5.50– 5.99	11
6.00– 6.49	10
6.50– 6.99	11
7.00– 7.49	8
7.50– 7.99	4
8.00– 8.49	4
8.50– 8.99	3
	269

The upper part of Fig. 35 indicates the method by which the ogive is built up. Just as in the histogram, the area of each rectangle is proportional to the number of cases falling in the given class. Since the operation is a cumulative one, however, the base of each rectangle is the cumulated frequencies of all preceding classes. Thus the y-value (frequency) of the first rectangle is 3, erected from zero as a base, the y-value of the second class is 10, erected from 3 as a base, and so on. The slope of the curve connecting these rectangles is gradual at first when the frequencies

[1] From "Labor Efficiency and Productiveness in Sawmills," Ethelbert Stewart, *Monthly Labor Review*, January, 1923, 14. Seven scattered cases above $9.00 in value have been omitted from the table and the accompanying graph.

are low, then steeper as the frequencies become greater, and finally tapers off as the frequencies decrease near the upper limit of the distribution. This is the cumulative frequency curve, or ogive.

When the various rectangles representing the class-frequencies are dropped to the zero line as a common base, the x-values remaining the same throughout, the histogram or column diagram described in an earlier section is secured. From this the frequency polygon or smoothed frequency curve may be derived.

REFERENCES

Bowley, A. L. *Elements of Statistics*, Book I, Chap. 4.

Chaddock, R. E. *Principles and Methods of Statistics*, Chaps. 4–5.

Croxton, F. E. and Cowden, D. J. *Practical Business Statistics*, Chap. 8.

Day, Edmund E. *Statistical Analysis*, Chaps. 1–8.

Jones, D. C. *A First Course in Statistics*, Chaps. 2–3.

Mudgett, Bruce D. *Statistical Tables and Graphs*.

Richardson, C. H. *An Introduction to Statistical Analysis*, Chap. 2.

Riggleman, J. R. and Frisbee, I. N. *Business Statistics*, Chap. 6.

Secrist, Horace. *Introduction to Statistical Methods*, Chap. 6.

Tippett, L. H. C. *The Methods of Statistics*, Chap. 1.

Waugh, A. E. *Elements of Statistical Method*, Chaps. 2–3.

Yule, G. U. and Kendall, M. G. *An Introduction to the Theory of Statistics*, Chap. 6.

CHAPTER IV

DESCRIPTION OF THE FREQUENCY DISTRIBUTION: AVERAGES

The classification of quantitative data and the construction of a frequency distribution constitute an important stage in the task of organization and analysis. By means of classification the underlying structure of the data may be revealed and the essential unity of a mass of material may be brought out. But this is only the first step in statistical analysis. It remains to develop methods of measuring and expressing more concisely the significant characteristics of a body of data. For certain purposes the frequency distribution itself must be summarized and condensed, must be boiled down until its essence has been distilled into three or four significant figures.

If each frequency distribution constituted a novel and unique problem, obeying a law peculiar to itself, the task of studying and describing such distributions would be a difficult one. Fortunately this is not so. Quantitative data in widely different fields, when assembled in frequency distributions, show certain common characteristics, obey certain general laws. Experience in one field, therefore, constitutes a guide to work in others. Uniformity in the behavior of masses of data makes possible the development of a generalized method of organizing, analyzing and comparing measurements drawn from many fields of scientific study.

Comparison of Frequency Distributions

This fact of a common law of arrangement running through the universe of quantitative facts may be brought home most effectively by a comparison of distributions illustrative

of various types of data. The characteristics of the frequency distributions and of the frequency curves which follow should be noted, and the distributions compared.

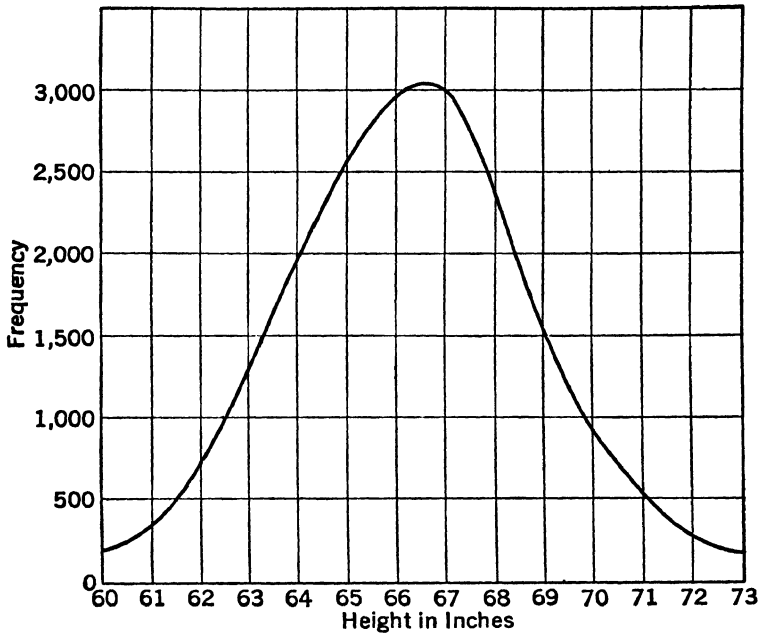


FIG. 37. — Frequency Curve: Distribution of 18,780 Soldiers Classified according to Height

The curve in Fig. 37 is based upon the following data relating to the heights of 18,780 soldiers.[1]

TABLE 16

Distribution of Soldiers Classified According to Height

Height in inches	*Number of soldiers*	*Height in inches*	*Number of soldiers*
60 +	197	67 +	3,017
61 +	317	68 +	2,287
62 +	692	69 +	1,599
63 +	1,289	70 +	878
64 +	1,961	71 +	520
65 +	2,613	72 +	262
66 +	2,974	73 +	174
	Total		18,780

[1] From G. C. Whipple, *Vital Statistics*, New York, Wiley, 1919, 377.

Fig. 38 depicts a frequency curve based upon 1,000 observations, made at Greenwich, of the Right Ascension of Polaris.[1] The values on the abscissa define deviations, in seconds of time, from an origin near the mean of all the observations. Frequencies of occurrence of given values on the x-scale are measured, of course, as ordinates on the

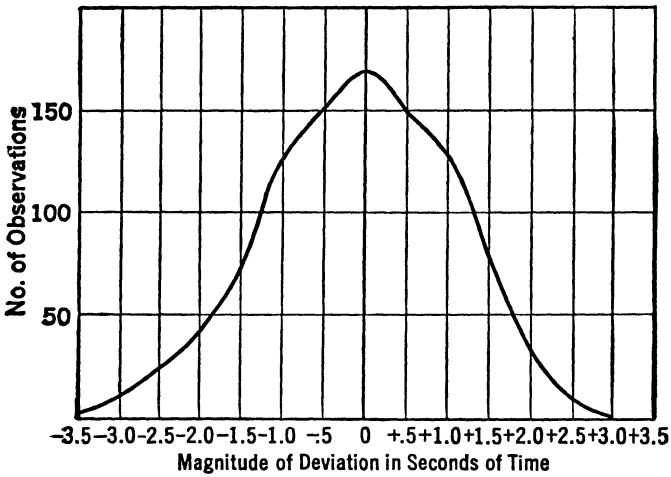


FIG. 38. — Frequency Curve: Distribution of Errors of Observation in Astronomical Measurements

y-scale. The distribution plotted in Fig. 38 is given in Table 17 on page 89.

If a piece of artillery be accurately adjusted on a given target (a point) and 100 shots be fired, it will be found that the points of impact of the hundred shots will be dispersed about the target. No matter how accurate the piece or the adjustment only a small percentage of the shots will fall upon the exact point at which they were directed. The points of impact will be scattered about the target in a quite regular fashion, however. If a rectangle be so drawn as to include all the points of impact, and this rec-

[1] From E. T. Whittaker and G. Robinson, *The Calculus of Observations*, London, Blackie and Son, 1924, 174.

TABLE 17

Distribution of Errors of Observation in Astronomical Measurements

(1000 observations of the Right Ascension of Polaris)

Magnitude of deviation, in seconds of time, from origin	*Number of observations*
− 3.5	2
− 3.0	12
− 2.5	25
− 2.0	43
− 1.5	74
− 1.0	126
− 0.5	150
0	168
0.5	148
1.0	129
1.5	78
2.0	33
2.5	10
3.0	2
	1,000

tangle (or *zone of dispersion*) be divided into eight equal parts, the distribution of shots within these sections will be as indicated in Fig. 39. (In any given case there are likely

2	7	16	25	25	16	7	2

FIG. 39. — Zone of Dispersion, Artillery Firing, Showing the Theoretical Percentage Distribution of Shots

to be slight departures from this order, but in the long run this distribution will prevail.)

This general rule holds for all classes of guns. The more accurate the gun the smaller will be the zone of dispersion, but the distribution within this zone is theoretically the same in all cases. Rules of fire used in artillery adjustment are based upon this fact.

The results of actual firing may be contrasted with this theoretical distribution. Table 18 presents a record of one thousand shots fired from a battery gun at the middle of a stationary target two hundred yards distant.[1] The target was divided by horizontal lines into eleven equal divisions.

TABLE 18

Distribution of One Thousand Shots from a Single Gun

Division	*Number of shots recorded*
1 (top)	1
2	4
3	10
4	89
5	190
6	212
7	204
8	193
9	79
10	16
11 (bottom)	2
	1,000

These results are presented graphically in Fig. 40.

The zone of dispersion being divided into eleven divisions instead of the eight referred to in describing the theoretical distribution, a direct comparison cannot be made. We have here, however, the same general type of distribution found in the other examples given. A tendency toward concentration in the lower half of the target reflects a slight departure from symmetry.

When coins are tossed the distribution of heads and tails is assumed to be determined by pure chance. In a single experiment ten coins were tossed 100 times. The following table shows the frequencies with which given numbers of heads appeared. (The greatest number of heads possible

[1] This experiment is recorded in the Report of the Chief of Ordnance, 1878, Appendix S. The results are given in *The Method of Least Squares*, Mansfield Merriman, New York, Wiley, 1897, 14.

in a given throw under such conditions is, of course, 10; it is also possible that no heads should appear.)

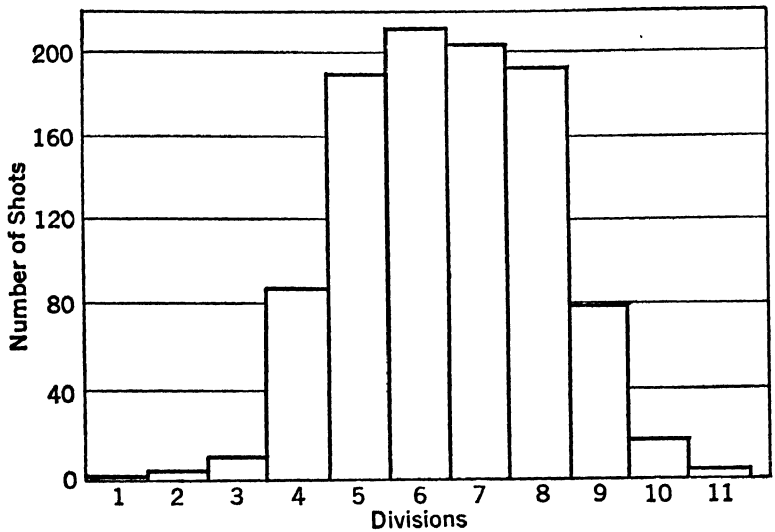


FIG. 40. — Column Diagram: Distribution of 1,000 Shots from a Single Gun

TABLE 19

Distribution of Results in Coin Tossing Experiment

(Ten coins tossed 100 times)

Number of heads	*Frequency of occurrence*
10	0
9	1
8	4
7	7
6	23
5	30
4	20
3	9
2	5
1	1
0	0
	100

Figure 41 depicts the above frequency distribution.

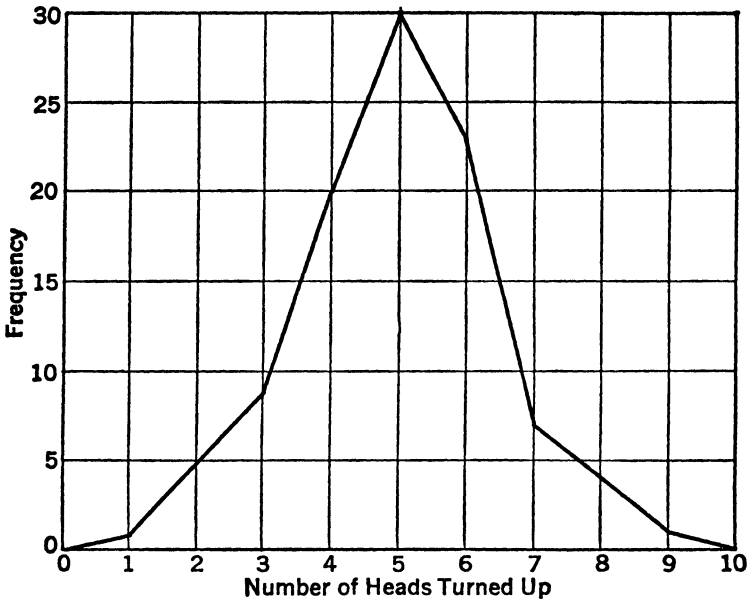


FIG. 41. — Frequency Polygon: Distribution of Heads in a Coin Tossing Experiment

DISTRIBUTION OF ECONOMIC DATA

We find in these four widely different fields something approaching a uniform law of arrangement of quantitative data. The examples which have been given, however, do not represent the world of economic facts. Do economic data show the same general characteristics? If reference be made to the examples given in Chapter III, comparisons with the four preceding illustrations may be made. The frequency distributions referred to are those relating to weekly earnings of employees, the length of life of telephone poles, the distribution of labor cost in sawmills and the distribution of incomes below $4,000 in the United States. (The curve of the latter distribution, it should be noted, would show a long tail extending far to the right if

the incomes above $4,000 were included.) Several additional examples of economic data may be given.

Figure 42 illustrates the order in which price variations are distributed. It is based upon a study made by W. C. Mitchell of 5,578 individual cases of change in the wholesale prices

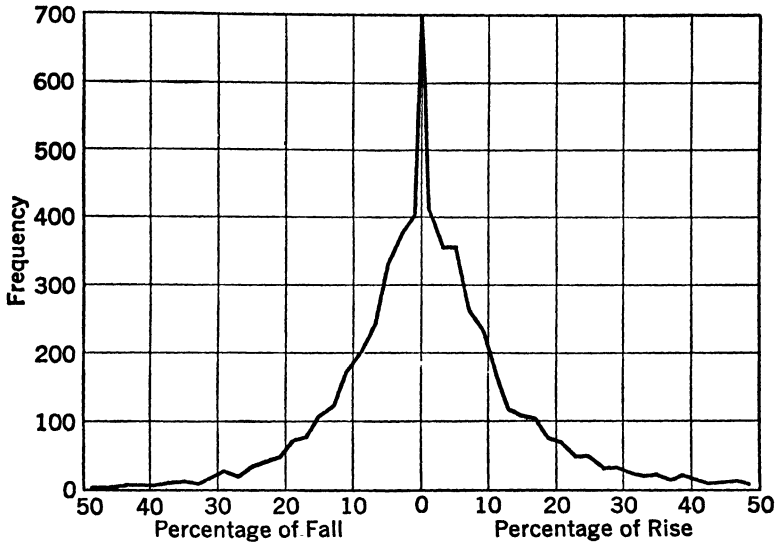


FIG. 42. — Frequency Polygon: Distribution of 5,540 Cases of Change in the Wholesale Prices of Commodities from One Year to the Next (after Mitchell)

of commodities from one year to the next.[1] Thus, for example, the average price of middling upland cotton in New York in a given year was $0.115 per pound. In the following year the average price was $0.128 per pound, an increase of 11.3 per cent. This would constitute one entry in the table of rising prices, falling in the class "10–11.9%." The entire table consists of 5,578 such entries. These data are presented in Fig. 42 in the form of a frequency polygon, no attempt being made to smooth the curve.

[1] From *Bulletin 284*, U. S. Bureau of Labor Statistics, Part I, "The Making and Using of Index Numbers," 18. The figure shows the price changes only within the range of a 51 per cent fall and a 51 per cent rise. One case of a price fall of 55 per cent is not shown, and 37 cases of price increases ranging from 52 per cent to 104 per cent have not been included.

Table 20 shows the distribution of London-New York exchange rates (sterling exchange) from 1882 to 1913, inclusive. This was a period when both currencies were freely convertible into gold, at fixed ratios, with customary market forces operating to keep exchange rates between the two "gold points." Observations covering recent decades would show quite different characteristics. In the distribution shown graphically in Fig. 43 monthly rates have been

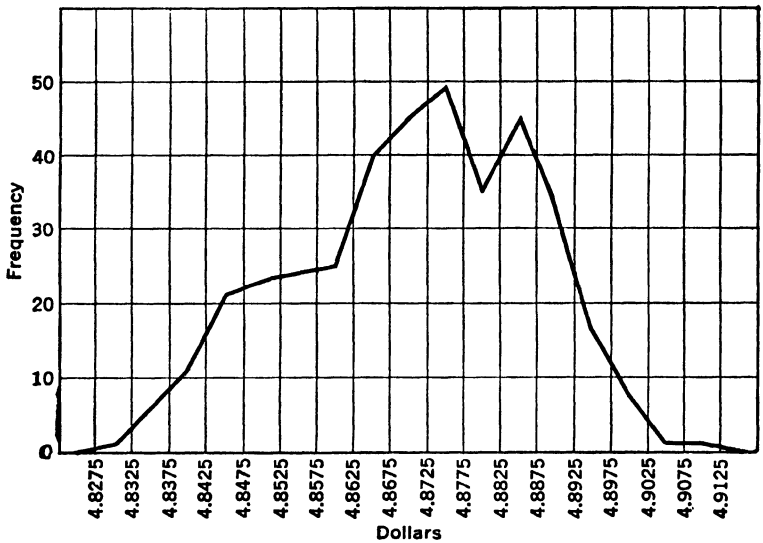


FIG. 43. — Frequency Polygon: Distribution of London-New York Exchange Rates (as recorded over a period of 384 months)

classified according to the frequency of their occurrence over thirty-two years of pre-war experience.[1]

A fairly typical distribution of wage-earners classified according to the amount of their weekly earnings, is shown in Table 21 and, graphically, in Fig. 44. The data relate to 13,427 steel workers in open-hearth furnaces, in the United

[1] "The figures are . . . the averages of those quoted at the beginning of each month in the *Economist;* on and after July, 1886, the exchange is the 'telegraphic transfer,' before that date, 'short at interest.' " The data are taken from *An Academic Study of Some Money Market and Other Statistics*, by E. G. Peake. London, P. S. King, 1923. Appendix I.

TABLE 20

Distribution of London-New York Exchange Rates as Recorded by Months during the Period 1882–1913

Class-interval	*Frequency (number of months given rate prevailed)*
$4.8275–$4.8324	1
4.8325– 4.8374	6
4.8375– 4.8424	11
4.8425– 4.8474	21
4.8475– 4.8524	23
4.8525– 4.8574	24
4.8575– 4.8624	25
4.8625– 4.8674	40
4.8675– 4.8724	45
4.8725– 4.8774	49
4.8775– 4.8824	35
4.8825– 4.8874	45
4.8875– 4.8924	33
4.8925– 4.8974	16
4.8975– 4.9024	8
4.9025– 4.9074	1
4.9075– 4.9124	1
	384

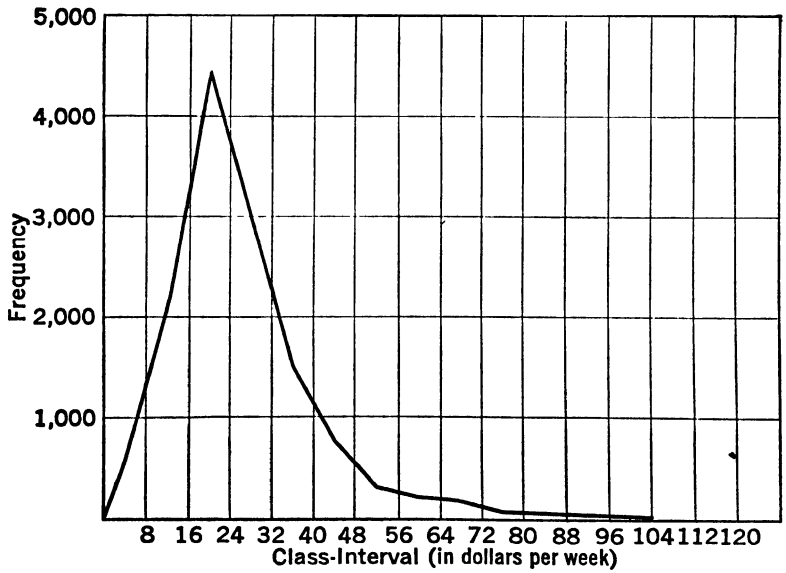


FIG. 44. — Distribution of Wage-Earners in Open-Hearth Furnaces, Classified according to Average Weekly Earnings in 1935

States, in 1935. There is a clear concentration of workers whose earnings fall between $16 and $24 a week. The distribution is markedly skewed, however, with a tail extending far to the right. The range of weekly earnings, like that of incomes in general, is far greater above the mode than below.

TABLE 21

Distribution of Wage-Earners in Open-Hearth Furnaces in the United States, Classified According to Average Weekly Earnings in 1935

(Total for all districts)

Class-interval (in dollars per week)	*Frequency (number of workers earning stated amount)*
$ 0– $ 7.99	583
8– 15.99	2,200
16– 23.99	4,462
24– 31.99	3,032
32– 39.99	1,527
40– 47.99	764
48– 55.99	358
56– 63.99	210
64– 71.99	144
72– 79.99	44
80– 87.99	36
88– 95.99	21
96– 103.99	26
104– 111.99	3
112– 119.99	7
120– 127.99	1
128– 135.99	9
	13,427

The frequency curves and histograms based upon economic data, it will be noted, do not all show the symmetry and regularity which seem to characterize the curves representing physical data. Some are non-symmetrical, showing a preponderance of cases on one side of the point of greatest concentration. In some there are breaks in the regularity of the increase or decrease of frequencies. But in spite of these differences there is obviously a family resemblance

between the measurements drawn from the fields of economics, astronomy, anthropometry, ballistics, and pure chance. Certain of the common characteristics may be noted.

General Characteristics of Frequency Distributions

There is, in the first place, *variation* in the values of the measurements secured. Human heights vary, astronomical measurements of the same quantity differ, projectiles fired under conditions as nearly constant as it is humanly possible to make them fail to land at the same spot, incomes vary as between individuals, and exchange rates move from week to week and month to month. The various observations or values secured in a given case are distributed along a scale, between two extreme values.

The distribution of these values along the scale (the x-axis) is such that, moving from one extreme value towards the other, the cases found at successive points along the scale (the successive class frequencies) increase with more or less regularity up to a maximum, and then decrease in much the same way. In spite of variation, therefore, we find a *central tendency*, a massing of cases at certain points on the scale of values. This is the second notable characteristic which all the frequency distributions appear to possess in common.

If we measure, for each of the successive classes, the amount of deviation along the scale from the point of greatest concentration it will be noted that small deviations are much more frequent than large ones, that extreme deviations are rare, and that deviations on both sides of the point of concentration reach perfect (or almost perfect) equality in the examples taken from the physical sciences and from the field of pure chance, and approximate equality in the economic distributions. (Exceptions to this rule of approximate equality on the two sides of the point of greatest concentration are not infrequent, the example of income distribution being a rather striking case in point.)

Figure 45 depicts a curve which is termed the "probability curve," or the "normal curve of error." Its characteristics will be discussed in greater detail in a later section. At this point it is presented merely as a basic type which some of the above examples approach closely, and from which others of the examples represent more or less pronounced deviations. Departures from this type, let it be emphasized, are numerous and significant, but as a basic

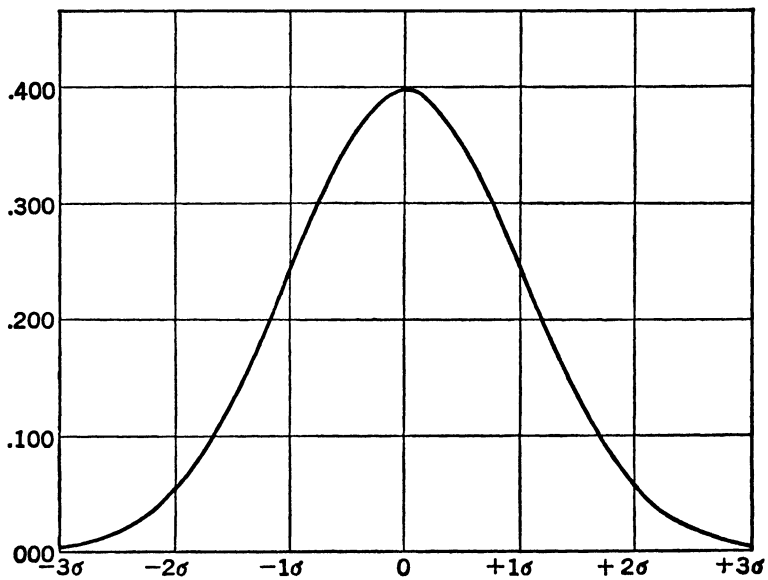


FIG. 45. — The Normal Curve of Error

form this normal curve of error is extremely important in statistical work. Even the most important variations from this type resemble it with sufficient closeness to justify the use of a generalized method of describing frequency distributions. Distributions of quantitative data vary, and their variations from each other and from certain standard types are of the greatest significance, but in spite of their variations a family resemblance runs through them all. Each new frequency distribution is not an isolated phenomenon,

but a member of a large family, and as such the problem of describing and analyzing it may be approached with confidence in methods which have been found applicable in other cases.

Given this more or less common type, how may a given distribution be described and differentiated from others? Certain methods will have been suggested by the preceding discussion.

Methods of Describing a Frequency Distribution

The values of all the observations, it has been noted, are spread along a scale. The frequency distribution may be described by the selection of a single value on that scale which is thoroughly representative of the distribution as a whole. Since the frequencies vary, an obvious choice is the selection of that value which occurs the greatest number of times, or, in other words, that point on the scale at which the concentration is greatest. This value constitutes a *measure of the central tendency* of the distribution. Thus, one might find the income class in which the greatest number of people fall, and let the mid-point of that class (which is $950 in the distribution presented in Table 11) serve as the representative of the distribution. This most common value, it should be noted, is only one of several possible measures of the central tendency of a given distribution. All such measures are termed *averages*.

A single representative value such as this has many uses but, by itself, it obviously leaves out many facts concerning the distribution. Of great importance is the character of the distribution about the average. Are the values of all tabulated cases closely concentrated, or is there pronounced dispersion over a wide range? The representative character of any average depends upon how closely the other values cling to it, upon the degree of concentration about the central tendency. The average, therefore, must be supplemented by a *measure of variation*, a measure of the "scatter" about the central value.

An adequate description should include also an account of the degree of symmetry of the distribution. It is highly important to know whether there is an equal distribution of cases on each side of the point of greatest concentration, or whether the frequency curve is skewed to one side, as in the case of income distribution illustrated above. If the curve is not symmetrical the degree of asymmetry should be determined, and for this purpose *measures of skewness* have been developed.

It is, finally, possible to measure the degree of peakedness of frequency curves, by comparing them with the normal curve of error as a standard. It is obvious that the frequency polygon representing price changes (Fig. 41) would, if smoothed, constitute a curve much more peaked than the normal curve, and this fact of pronounced concentration at the central value is highly significant. This characteristic of frequency curves is called *kurtosis*, and the measurement of kurtosis constitutes the final step in the description of the frequency distribution.

When these various measures have been secured the task of statistical analysis will be well under way. The chaotic assortment of data with which we started will have been reduced to workable form in the shape of a frequency table, and the essential facts which the table reveals will have been distilled into three or four significant measures. This process not only reveals the characteristics of the given distribution, but also facilitates comparison with similar distributions. For example, it is impossible to compare some tens of millions of unorganized personal income figures for the United States with similar data for Great Britain. But if we secure a value for the average or most representative income for each country, together with a description of the distribution of personal incomes about that central value, a legitimate basis for comparative study is obtained. In manipulating and analyzing masses of material, whatever the purpose of study may be, full use

should be made of the power to condense, simplify and compare which is given by the measures employed in describing the frequency distribution.

The succeeding section is devoted to a discussion of one phase of this descriptive process, that concerned with the measurement of central tendencies. After the development of this subject of averages, problems relating to measures of variation and of skewness will be dealt with.

Averages

We have seen that the representation of a frequency distribution by an average, a single typical figure, is justified because of the tendency of large masses of figures to cluster about a central value, from which the values of all observed cases depart with more or less regularity and smoothness. It is solely because of the concentration of cases about a central point on the scale that such representative figures have significance. The average represents the distribution as a whole only because it is a typical value. If the individual items entering into a distribution vary widely in value and show no tendency toward concentration, no single value can represent them. Thus the arithmetic mean of the three numbers 3, 125, 1,000 is 376, but 376 in no way represents the three values on which it is based. This fundamental requirement, that there be a tendency toward concentration about a central value, must be met if an average is to be at all representative.

If the general character of a frequency distribution be recalled the logic of one sort of average will be clear at once. It was suggested above that that point on the x-scale at which the concentration is greatest, that value which occurs the greatest number of times, might be taken as typical of the entire distribution. This value is termed the *mode*, and the group in which it falls is called the *modal group*. If a frequency curve be drawn to represent a given distribution, the mode will be the *x-value corresponding to*

the maximum ordinate.[1] The maximum ordinate itself measures the frequency of the modal group. Students frequently confuse these two values in determining the mode. It is not the distance along the y-scale but the distance along the x-scale which measures the value of the mode. The ordinates merely measure the number of cases falling in the several classes, not the values of the cases falling in those classes.

As typical of a given distribution we might also select that point on the scale of x-values on each side of which one half the total number of cases fall. This value, which is called the *median,* is that which exceeds the values of one half the cases included, and is in turn exceeded by the values of one half the cases. Thus it has been estimated that in 1918 the median value of personal incomes in the United States was $1,140; one half of the 37 million recipients of personal incomes received less than this sum, while one half received more. When a distribution is represented by a frequency curve, the area under the curve is divided into two equal parts by an ordinate erected at that point on the x-axis corresponding to the median value. This follows, of course, from the definition of the median, and from the fact that the area under a frequency curve represents the total number of cases included in the distribution.

The *arithmetic mean* is a third type of average which may be used to represent a distribution. This is a *calculated* average, affected by the value of every item in the distribution. Herein, obviously, it differs from the mode and the median, which depend primarily upon the relative position of the items in the frequency table, and are not affected by the values of all individual items. The arithmetic mean is the center of gravity of a distribution; it would be the x-value of the point of balance of a frequency

[1] Strictly speaking, the mode is the x-value corresponding to the maximum ordinate of the ideal frequency curve which has been fitted to the given distribution.

curve, if the curve could be blocked out and manipulated in solid form.

The *geometric mean* and the *harmonic mean* are two other averages the characteristics of which will be discussed at a later point.

The computation or location of these various averages may involve somewhat lengthy processes if the number of cases included be great. If appropriate methods be employed, however, the labor of computation may be materially cut down. The use of the following symbols will simplify the explanation of these methods:

M: Arithmetic mean.
Mo: Mode.
Md: Median.
m: The value of an individual observation; in a frequency distribution, the value of the midpoint of a class.
f: The number of items (observations) in a given class in a frequency distribution.
N: The total number of items in a given series or frequency distribution.
Σ (Sigma): The symbol for the process of summation, meaning "the sum of."

The Computation of the Arithmetic Mean

Using the above notation, the formula for the arithmetic mean is

$$M = \frac{\Sigma m}{N}.$$

Thus the mean of the measures 2, 5, 6, 7, is equal to the sum of these measures divided by 4, which is $\frac{20}{4}$ or 5. The computation of the arithmetic mean when each measure is reported at its true value is thus a simple process of summation and division. The weekly earnings of 210 factory employees were listed in an earlier section. If these figures be added, and the total divided by 210, the mean weekly

wage is found to be \$26.983. In this case the task of adding 210 items is somewhat tedious; it is a task which would become almost impossible if one were dealing with the 37 million personal income figures, for example. For practical reasons, therefore, it is usually necessary to compute the required averages from the frequency distribution rather than from the original ungrouped data. To exemplify this process we may utilize data relating to the weekly earnings of steel workers in the Pittsburgh District in 1935.

The importance of certain of the precautions mentioned in the section on classification, in connection with the choice of a class-interval, will be clear from this example. When the mean of a distribution is calculated from classified observations, we must assume an even distribution of cases within each class. The class-interval should be selected with this in mind, in order that errors introduced by the assumption may be minimized. If the items in each class are evenly distributed, the mid-value of each class may be taken as representative of all the observations included; when such a mid-value is multiplied by the number of items in the class, the product is approximately equal to the sum of all the individual items in the class. The formula for the mean thus becomes $M = \frac{\Sigma(fm)}{N}$. Table 22 illustrates the procedure in detail.

The value secured in this way is sometimes called a weighted arithmetic mean. What we do, in effect, is to secure the arithmetic mean of the 28 figures in the column headed *m*. We do not take a simple average of these figures, however, but *weight* each one in proportion to the number of cases falling in the class-interval of which it is the mid-value. It is precisely the procedure we should follow in calculating the mean of five men's incomes, two of whom, let us say, have incomes of \$2,000 and three of whom have incomes of \$3,000. Clearly it would not do to add the figures \$2,000 and \$3,000, dividing the sum by two. The

TABLE 22[1]

Calculation of the Arithmetic Mean of Weekly Earnings of Workers in Open-Hearth Furnaces in the Pittsburgh District in 1935

Class-interval (*in dollars per week*)	*Mid-point* m	*Frequency* f	fm
$ 0–$ 3.99	2	67	134
4– 7.99	6	290	1,740
8– 11.99	10	437	4,370
12– 15.99	14	730	10,220
16– 19.99	18	1,056	19,008
20– 23.99	22	1,009	22,198
24– 27.99	26	712	18,512
28– 31.99	30	609	18,270
32– 35.99	34	334	11,356
36– 39.99	38	187	7,106
40– 43.99	42	179	7,518
44– 47.99	46	105	4,830
48– 51.99	50	60	3,000
52– 55.99	54	67	3,618
56– 59.99	58	28	1,624
60– 63.99	62	37	2,294
64– 67.99	66	33	2,178
68– 71.99	70	29	2,030
72– 75.99	74	16	1,184
76– 79 99	78	8	624
80– 83.99	82	3	246
84– 87.99	86	8	688
88– 91.99	90	4	360
92– 95.99	94	7	658
96– 99.99	98	9	882
100–103.99	102	5	510
104–107.99	106	1	106
108–111.99	110	1	110
Total		6,031	145,374

$$M = \frac{\Sigma(fm)}{N} = \frac{\$145,374}{6,031} = \$24.1045.$$

[1] These figures and similar data appearing in subsequent tables were compiled by Edward K. Frazier, of the Division of Wages, Hours and Working Conditions, U. S. Bureau of Labor Statistics. See "Earnings and Hours in Blast Furnaces, Bessemer Converters, Open-Hearth Furnaces and Electric Furnaces, 1933 and 1935" *Monthly Labor Review*, April, 1936. The detailed statistics in Table 22 were provided through the courtesy of Dr. Isador Lubin Commissioner of Labor Statistics.

figure $2,000 is given a weight of two, the figure $3,000 is given a weight of three, and the resultant sum, $13,000, is divided by five. Though the procedure in working from the frequency distribution is thus a form of weighting, the term "weighted average" is coming to have a more restricted meaning, to be explained at a later point, and should not in general be applied to an average computed from a frequency distribution.

SHORT METHOD OF COMPUTING THE ARITHMETIC MEAN

The calculation of the arithmetic mean from the frequency table is much easier, in general, than from the ungrouped data, but when the number of cases included is large even the computation from the frequency table by the method illustrated above may be laborious. The procedure may be greatly simplified.

From the method of computing the arithmetic mean it follows that the algebraic sum of the deviations of a series of individual magnitudes from their mean is zero. This may be readily demonstrated. We may represent the series of magnitudes by m_1, m_2, m_3, . . . m_n, their arithmetic mean by M, and the deviations of the various magnitudes from the mean by d_1, d_2, d_3, . . . d_n.

Then

$$\frac{m_1 + m_2 + m_3 + \ldots + m_n}{N} = M \qquad (1)$$

and

$$m_1 + m_2 + m_3 + \ldots + m_n = NM. \qquad (2)$$

The number of terms, of course, is equal to N. Therefore, subtracting M N times from each side of the equation,

$$(m_1 - M) + (m_2 - M) + (m_3 - M) + \ldots + (m_n - M) = 0. \qquad (3)$$

But

$m_1 - M = d_1$, $m_2 - M = d_2$, etc., and equation (3) may be written

$$\Sigma d = 0.$$

Knowing this to be true we may measure the deviations of a series of magnitudes from any arbitrary quantity,

secure the algebraic sum of the deviations, and from this value ascertain the difference between the arbitrary quantity and the true mean. For this difference will be the mean of the deviations from the arbitrary origin. If we let M' represent the arbitrary origin, or assumed mean, while $c = M - M'$, and d_1', d_2', d_3' . . . d_n', represent the deviations of the various magnitudes from M' (i.e., $d_1' = m_1 - M'$, $d_2' = m_2 - M'$, etc.), then

$$d_1' = d_1 + c,\ d_2' = d_2 + c,\ d_3' = d_3 + c,\ \ldots\ d_n' = d_n + c$$

and

$$\Sigma d' = \Sigma d + Nc.$$

But

$$\Sigma d = 0$$

$$\therefore \Sigma d' = Nc$$

and

$$c = \frac{\Sigma d'}{N}.$$

From the known values of M' and c the value of the true mean may be obtained, for $M = M' + c$. The procedure is illustrated in the following simple example:

TABLE 23

Computation of the Arithmetic Mean (Short Method)

(Ungrouped data)

m	f	d'
5	1	− 15
15	1	− 5
25	1	+ 5
35	1	+ 15
45	1	+ 25
	5	+ 25

$M' = 20$

$$c = \frac{\Sigma d'}{N} = \frac{+25}{5} = 5$$

$$M = M' + c = 20 + 5 = 25$$

When the deviations are measured from 20 as arbitrary origin there is in each case a constant error, if the deviation from the true mean be taken as standard. This error is equal to the difference between the true and the assumed means. The algebraic sum of the deviations from the assumed mean will equal N times this constant error, since

the error is repeated once for every item included. By dividing the sum of these deviations by N the amount of the error may be determined and the value of the mean thus obtained.

TABLE 24

Calculation of the Arithmetic Mean of Weekly Earnings of Workers in Open-Hearth Furnaces in the Pittsburgh District in 1935

(Short method)

Class-interval (in dollars per week)	*Mid-point* m	*Frequency* f	d' *(in class-interval units)*	fd' −	fd' +
$ 0–$ 3.99	2	67	− 7	469	
4– 7.99	6	290	− 6	1,740	
8– 11.99	10	437	− 5	2,185	
12– 15.99	14	730	− 4	2,920	
16– 19.99	18	1,056	− 3	3,168	
20– 23.99	22	1,009	− 2	2,018	
24– 27.99	26	712	− 1	712	
28– 31.99	30	609	0		
32– 35.99	34	334	+ 1		334
36– 39.99	38	187	+ 2		374
40– 43.99	42	179	+ 3		537
44– 47.99	46	105	+ 4		420
48– 51.99	50	60	+ 5		300
52– 55.99	54	67	+ 6		402
56– 59.99	58	28	+ 7		196
60– 63.99	62	37	+ 8		296
64– 67.99	66	33	+ 9		297
68– 71.99	70	29	+ 10		290
72– 75.99	74	16	+ 11		176
76– 79.99	78	8	+ 12		96
80– 83.99	82	3	+ 13		39
84– 87.99	86	8	+ 14		112
88– 91.99	90	4	+ 15		60
92– 95.99	94	7	+ 16		112
96– 99.99	98	9	+ 17		153
100–103.99	102	5	+ 18		90
104–107.99	106	1	+ 19		19
108–111.99	110	1	+ 20		20
Total		6,031		− 13,212	+ 4,323

Calculations $M' = \$30$

1. Algebraic sum of deviations from M'

   − 13,212
   + 4,323
   − 8,889

2. Calculation of c (in class-interval units)

$$c = \frac{-8{,}889}{6{,}031} = -1.47388$$

3. Reduction of c to original units

   Class-interval = \$4
   c (in original units)
   $= -1.47388 \times \$4$
   $= \$ - 5.8955$

4. Determination of M

   $M = M' + c$
   $M = \$30 - \$5.8955$
   $M = \$24.1045$ ✓

The work of computation may be still further abbreviated, for observations arranged in the form of a frequency distribution, by measuring the deviations in terms of the class-interval as a unit. Then, in finally applying the necessary correction, the difference between the true and assumed means may be again expressed in terms of the original units.

The method may be illustrated in detail with reference to the wage data for which the mean has already been calculated.

The steps in this process of calculating the arithmetic mean by the short method may be briefly summarized:

1. Organize the data in the form of a frequency distribution.
2. Adopt as the assumed mean the midpoint of a class near the center of the distribution.
3. Arrange a column showing the deviation (d') from the assumed mean of the items in each class, in terms of class-interval units. This deviation will be zero for the items in the class containing the assumed mean, -1 for the items in the next lower class, $+1$ for the items in the next higher class, and so on.
4. Multiply the deviation of each class by the frequency of that class, taking account of signs. These products are entered in the column fd'.
5. Get the algebraic sum of the items entered in the column fd'.
6. Divide this sum by the total frequency (N). The quotient is the correction (c) in class-interval units.
7. Multiply the correction (c) by an amount equal to the class-interval. The product is the correction in terms of the original units.
8. Add this correction (algebraically) to the assumed mean (M'); the sum is the true mean (M).

Location of the Median

UNGROUPED DATA

The median is a value of a variable so selected that 50 per cent of the total number of cases, when arranged in order of magnitude, lie below it and 50 per cent above it. For many frequency distributions this is a useful and significant value.

When handling data which are not arranged in the form of a frequency distribution the location of the median is a simple matter. The data having been arranged in order of magnitude, it is necessary only to count from one end until that point on the scale of values is found which divides

the number of cases into two equal parts. As a simple example we may assume that the following seven figures represent the annual incomes of seven individuals:

$750 $975 $1,128 $1,450 $1,475 $1,825 $1,950

The scale of values extends from $750 to $1,950, and seven items are arranged along this scale. The value of $1,000 has two items on one side and five items on the other, so obviously does not conform to our definition of

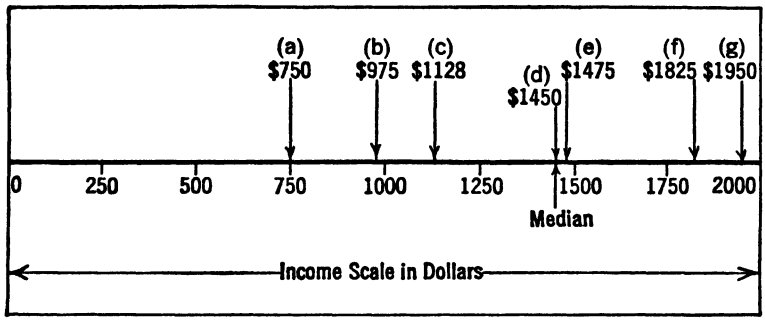


FIG. 46. — Illustrating the Location of the Median with Ungrouped Data (Personal incomes of seven individuals)

the median. The value of $1,450, which corresponds with the income of one of the seven individuals, is the median in this case. Three items lie on each side of this value; or, if we assume the central item to be cut in two, $3\frac{1}{2}$ items lie on each side of this point. This case is illustrated in Fig. 46. This diagram may help to bring out the fact that the median is a point on a scale so located that it cuts the frequencies in two.

The problem is slightly different when an even number of cases is included. This condition is exemplified in the table on page 111 which shows the average earnings per man-hour in each of 38 selected industries during the year 1933.

In this case the median must be a value on each side of which 19 industries lie. Therefore any value exceeding $0.425 (average earnings in the prepared feed industry) and less than $0.426 (average earnings in the meat packing

TABLE 25

Average Earnings per Man-Hour in Selected Manufacturing Industries [1]

	Average wage per man-hour
Silk and rayon goods: Commission throwing	$.278
Cotton goods	.279
Cigars	.299
Silk and rayon goods: Commission weaving	.313
Silk and rayon goods: Regular throwing	.316
Knit underwear	.319
Knit outerwear	.358
Cigarettes	.361
Silk and rayon goods: Regular weaving	.369
Wool shoddy	.370
Hosiery	.372
Cotton small wares	.378
Woolen goods	.395
Sugar, beet	.395
Worsted goods	.399
Snuff, and chewing and smoking tobacco	.402
Knit cloth	.414
Rayon yarns	.421
Feeds, prepared	.425
Meat packing	.426
Pulp	.431
Ice, manufactured	.436
Flour milling	.444
Paper	.445
Carpets and rugs, wool	.464
Leather tanning	.470
Sugar refining, cane	.481
Soap	.482
Blast furnaces	.488
Felt goods	.488
Cereal preparations	.510
Steel works	.519
Motor vehicle bodies and parts	.561
Machine tools	.585
Motor vehicles	.610
Machine-tool accessories	.621
Petroleum refining	.643
Malt	.657

[1] From *Monthly Labor Review*, October, 1935, 910.

industry) will satisfy the definition of a median. Under these conditions, where the median is really indeterminate, a value half-way between the two limiting values is accepted, by convention. The median of the 38 figures would thus be $0.4255.

In this example the median value does not correspond with the earnings in any one industry. This will frequently be so when there is an even number of observations.

GROUPED DATA

The task of locating the median is essentially the same when the data are in the form of a frequency distribution. The fact that the real values of the individual items are not known, because of the grouping by classes, complicates the problem slightly. The data in Table 26, relating to advertising rates of daily newspapers in the United States, may be used in illustrating the method.

TABLE 26

Location of Median, Newspaper Advertising Rates in 1933 *Minimum Line Rates for National Advertising,* 245 *Daily Newspapers in Cities of* 25,000 *to* 50,000 *Population* [1]

Class-interval Rate per line (in cents)	*No. of newspapers charging stated rate* f
1.0– 2.99	6
3.0– 4.99	53
5.0– 6.99	85
7.0– 8.99	56
9.0–10.99	21
11.0–12.99	16
13.0–14.99	4
15.0–16 99	4
	245

$$\frac{N}{2} = \frac{245}{2} = 122.5$$

$$Md = 5.0 + \left(\frac{63.5}{85} \times 2.0\right)$$

$$= 5.0 + 1.49$$

$$= 6.49$$

In the present case the location of the median involves the determination of that value on each side of which 122.5

[1] Source: Editor and Publisher, *International Yearbook* for 1933.

items lie. We may assume that we start at the lower end of the scale and move through the successive classes. When we reach the upper limit of the first class (that including items having values from 1.0 to 3.0) we have left behind us 6 cases, while 239 lie in front of us. When the upper limit of the second class is attained, 59 items have been passed. The upper limit of the third class has below it 144 items. Somewhere between the lower and upper limits of the third class lies the desired point, that which has 122.5 items on each side of it. How far must we move through this class, from 5.0 to 7.0 in order to reach this point?

It will be recalled that, for purposes of calculation, the assumption is made that there is a uniform distribution of the items lying within any given class. Since before we reach the third class 59 cases have been counted, only 63.5 of the 85 included in this class are needed to complete the desired number, 122.5. On the assumption of even distribution the required 63.5 cases will lie within a distance on the scale equal to $\frac{63.5}{85}$ of the class-interval. The class-interval is 2.0; $\frac{63.5}{85}$ of 2.0 is equal to 1.49. As we move up the scale, then, having reached 5.0, we proceed an additional distance equal to 1.49. At a point on the scale having a value of 6.49 is the dividing line on each side of which lie 122.5 cases. This is the value of the median.

The process of computation is shown at the right of the frequency table. The following is a summary of the steps involved in the location of the median:

1. Arrange the data in the form of a frequency distribution.
2. Divide the total number of measures by 2; this gives the number which must lie on each side of the point to be located.
3. Begin at the lower end of the scale and add together the frequencies in the successive classes until the lower limit of the class containing the median value is reached.

4. Determine the number of measures from this class which must be added to the frequencies already totaled to give a number equal to $N/2$.
5. Divide the additional number thus required by the total number of cases in the class containing the median. This indicates the fractional part of the class-interval within which the required cases lie.
6. Multiply the class-interval by the fraction thus set up.
7. To the lower limit of the interval containing the median add the result of the multiplication process indicated in (6). This gives the value of the median.

The last three steps constitute merely a simple form of interpolation.

The entire process may be reversed by beginning at the upper end of the scale and counting downwards. In this case the final operation is one of subtraction from the upper limit of the interval containing the median.

N/2 may be a fractional value, as in the example given, or a whole number. The operation is precisely the same in the two cases.

Quartiles and Deciles

For many purposes it is desirable to locate on the scale of values, along which the items constituting a frequency distribution are ranged, points dividing the total number of measures in other ways. Similar to the median, which divides the total number of cases into two equal groups, are the quartiles, deciles, and percentiles. The quartiles, as the term implies, are points on the scale which divide the entire number of measures into four equal groups, the deciles divide the number into ten equal groups, and the percentiles divide the total number of cases into 100 equal groups. Thus the first quartile is that point on the scale below which one quarter of the total number of cases lie and above which three quarters of the total number of cases lie. The second quartile and the median are identical values. The third decile is that point on the scale below

which three tenths of the total number of cases lie and above which seven tenths of the total number of cases lie. In all cases the count begins at the lower end of the scale.

Example: Location of the First Quartile (Q_1), Newspaper Advertising Rates (See Table 26)

$$N/4 = 61.25$$
$$Q_1 = 5.0 + (2.25/85 \times 2.0)$$
$$= 5.05$$

Example: Location of Eighth Decile (D_8), Newspaper Advertising Rates (See Table 26)

$$N/10 = 24.5 \qquad D_8 = 7.0 + (52/56 \times 2.0)$$
$$8N/10 = 196 \qquad = 8.86$$

A method of locating median, quartiles, deciles and percentiles graphically is explained below.

Location of the Mode

The mode is the value of the x-variable corresponding to the maximum ordinate of a given frequency curve. The concept of a modal value is a thoroughly easy one to grasp. It is the most common wage, the most common income, the most common height. It is the point where the concentration is greatest, a characteristic which is effectively brought out by Fechner's term for this average, *dichtester wert*, or *thickest value*. It is not so easy, however, to locate the true modal value in a given case. In general statistical work an approximate value only is secured for the mode, but for most practical purposes this value is usually sufficiently accurate.[1]

The method of determining this approximate modal value may be illustrated by reference to the distribution shown in Table 27 on page 116.

There is wide dispersion of the 22 cases falling below 40, and the existence of this "open-end" class makes it impossible to compute the mean, as the table stands. The mode

[1] A method of locating the mode more accurately is explained in a later section.

TABLE 27

Frequency Distribution of Five Per Cent Bonds

(This table is based upon quotations on the New York Stock Exchange on June 13, 1936, on railroad and industrial bonds with coupon rate of 5 per cent)

Quoted price *Class-interval*	*Mid-point* *m*	*Frequency* *f*
Less than 40		22
40– 49.9	45	5
50– 59.9	55	5
60– 69.9	65	3
70– 79.9	75	8
80– 89.9	85	9
90– 99.9	95	19
100–109.9	105	49
110–119.9	115	10
120–129.9	125	3
130–139.9	135	1
		134

is therefore an appropriate average to employ in the present instance.

The class having limits of 100–109.9 contains the greatest number of cases. This appears to be the modal group, and the mid-point of this class, 105, may be tentatively accepted as the value of the approximate mode. But with different classifications quite different values might be secured for the mode. When the original bond quotations are tabulated with varying class-intervals the following results are secured. (Only the frequencies of the central classes are shown. It is not necessary, for this purpose, to present each of the tables as a whole.)

(*a*)		(*b*)		(*c*)		(*d*)	
Class-interval = 5		*Class-interval* = 2.5		*Class-interval* = 2.5		*Class-interval* = 1	
Class-interval	*f*	*Class-interval*	*f*	*Class-interval*	*f*	*Class-interval*	*f*
80– 84.9	3	90.0– 92.49	4	98.75–101.249	6	100–100.9	1
85– 89.9	6	92.5– 94.99	6	101.25–103.749	17	101–101.9	2
90– 94.9	10	95.0– 97.49	2	103.75–106.249	20	102–102.9	9
95– 99.9	9	97.5– 99.99	7	106.25–108.749	8	103–103.9	10
100–104.9	29	100.0–102.49	9			104–104.9	7
105–109.9	20	102.5–104.99	20			105–105.9	6
110–114.9	7	105.0–107.49	13			106–106.9	5
115–119.9	3	107.5–109.99	7			107–107.9	4

With a class-interval of 5 a value of 102.5 is secured for the mode; with a class-interval of 2.5 a value of 103.75 is obtained. A class-interval of 2.5, again, but with different class limits, yields a mode of 105. Finally, a class-interval of 1 gives a mode of 103.5. Further changes in classification would give still other values. The mode thus appears to be a curiously intangible and shifting average. Its value, for the same data, seems to vary with changes in the size of the class-interval and in the location of the class-limits.

These difficulties arise primarily from limitations to the size of the sample being studied. The true mode, that value which would occur the greatest number of times in an infinitely large sample, could be located exactly if we could increase indefinitely the number of cases included. For, given sufficient cases, the approximate mode approaches the true mode as the class-interval decreases. Grouping in large classes obscures details, and as these classes are reduced in size more of the details are seen and a truer picture of the actual distribution is secured. But since most practical work is necessarily based upon relatively small samples, the increase in the number of classes reveals gaps and irregularities, and causes such a loss of symmetry and order that doubt arises as to where the point of greatest concentration really lies. The different tabulations of bond prices furnish an excellent example of this.

By mathematical methods it is possible to obtain a value for the true mode without securing an infinite number of cases. The smoothing process has been briefly explained. One sort of smoothing involves the fitting of an appropriate type of ideal frequency curve to the data of a given frequency distribution. This gives, theoretically, the distribution which would be secured by the process first indicated, that of decreasing indefinitely the size of the class-interval and increasing indefinitely the number of cases. The value of the x-variable corresponding to the maximum ordinate of this ideal fitted curve is the true mode.

For most practical purposes approximate values of the mode are adequate, and these may be secured by much simpler methods. A first and rough approximation may be obtained by taking the mid-value of the class of greatest frequency, a method suggested above. If the general rules for classification which were outlined in an earlier section have been followed, this procedure will not generally involve a gross error.

It is possible, given a fairly regular distribution, to secure, by a process of interpolation within the modal group, a closer approximation than is obtained by accepting the mid-value of this group as the mode. Referring again to the tabulation of bond prices in Table 27 it will be noted that the distribution on the two sides of the modal class is not symmetrical. The modal class is that with a mid-value of 105. The class next below, with a mid-value of 95, contains 19 cases, while that next above, with a mid-value of 115, contains but 10 cases. The disproportion is continued in the succeeding classes below and above, more cases being bulked below the modal class than above. For other purposes we have assumed an even distribution of cases between the upper and lower limits of each class, but it is probable that this is not true of the modal class in the present case. Judging from the distribution outside this class, it is likely that the concentration is greater in the lower half of the class-interval, that is, between 100 and 105. The mode, therefore, probably lies below the mid-value 105, rather than precisely at that point. We may attempt to locate it within the group by weighting, assuming a pull toward the lower end of the scale equal to 19 (the number in the class next below) and a pull toward the upper end of the scale equal to 10 (the number in the class next above). This may be expressed by a formula, employing the following symbols:

l = lower limit of modal class.
f_1 = frequency of class next below modal class in value.
f_2 = frequency of class next above modal class in value.
i = class-interval.

The interpolation formula is

$$Mo = l + \frac{f_2}{f_2 + f_1} \times i.$$

Applying this formula to the bond price data presented in Table 27, we have

$$Mo = 100 + \left(\frac{10}{29} \times 10\right) = 100 + 3.45 = 103.45.$$

A closer approximation may sometimes be secured by basing the weights (represented by f_2 and f_1) upon the total frequencies of the two or three classes next above the modal class and the same number below. If three classes on each side are included in the present case, a value of 102.8 is secured for the mode of bond prices.

In some cases the problem of locating the mode is complicated by the existence of several points of concentration, rather than the single point which has been assumed in the preceding explanation. Thus in Table 9, representing the distribution of wages, with a class-interval of 25 cents, there are two definite modal points. A distribution of this type is called bi-modal; when plotted, a frequency curve having two humps is obtained. If the data are homogeneous such a distribution is the result of paucity of data and of the method of classification employed. It may be due to the use of a class-interval too small, with respect to the number of cases included in the sample. An approximate mode may be determined in such cases by shifting the class-limits and increasing the class-interval, carrying on this process until one modal group is definitely established. This reverses the process by which the true mode may be located when the number of cases is infinitely large. Under such conditions the class-interval might be reduced until it was infinitely small. But with a limited number of cases the location of the point where the concentration is greatest necessitates increasing the size of the class-interval, in order

to get away from the irregularities due to the smallness of the sample.

If the distribution remains bi-modal in spite of changes in the class-intervals and class-limits, it is probable that the data are not homogeneous, that two different distributions have by mistake been combined. Such cases are not uncommon in biometrical work. The existence of two distinct animal species where only one was suspected has been revealed in this way. The whole significance of a frequency distribution will be lost if the data are not homogeneous, a fact which is as true of work in the field of economic statistics as in any other.

DETERMINATION OF THE MODAL VALUE FROM MEAN AND MEDIAN

Another method of securing an approximate value for the mode, a method based upon the relationship between the values of the mean, median and mode, may be employed in certain cases. In a perfectly symmetrical distribution mean, median and mode coincide. As the distribution departs from symmetry these three points on the scale are pulled apart. If the degree of asymmetry is only moderate the three points have a fairly constant relation. The mode and mean lie farthest apart, with the median one third of the distance from the mean towards the mode. If the asymmetry is marked, no such relationship may prevail. Having the values of any two of the averages in a moderately asymmetrical frequency distribution, therefore, the other may be approximated. In fact, however, the method should only be employed in determining the value of the mode, as the other two values may be computed more accurately by other methods. The value of the mode itself should only be determined in this way when more exact methods are not applicable or are not called for.

The following formula is based upon this relationship:

$$Mo = \text{Mean} - 3(\text{Mean} - Md).$$

Applying this formula to the telephone pole data shown in Table 12, the following result is secured:

$$Mo = 9.33 - 3(9.33 - 9.015) = 8.385.$$

This value is slightly below the mid-value of the modal class, 8.5, and is also less than the value 8.49 which is secured by weighting within the modal group (using four classes on each side).

It must be emphasized that there is a fictitious accuracy to all these values for the mode. All the methods of locating the mode which have been discussed are merely approximative, a fact not to be forgotten in interpreting and utilizing the results.

Graphic Location of Mode, Median, Quartiles, and Deciles

A better understanding of the frequency curve and of the cumulative frequency curve may be secured through a brief discussion of certain methods of locating graphically some of the statistical measures that have been described.

The value of the mode may be readily determined from a frequency curve of the usual type, for, by definition, the mode is the reading on the horizontal scale corresponding to the maximum ordinate of such a curve. If this reading be taken from the frequency polygon a rough value will be obtained, the mid-value of the class of greatest frequency. A closer approximation to the true value of the mode will be secured from a curve which has been smoothed, either by inspection or by mathematical methods. Figure 47, showing a curve (smoothed by inspection) based upon the wage data presented in Table 8, indicates how the mode may be located graphically. The horizontal reading corresponding to the maximum ordinate of this curve is $27.50, an approximate value of the mode which may be compared with the values of $27.69 secured by the weighting process

and of $27.3470 secured from the values of the mean and median.

The locations of the median and mean have been indicated on this chart. It has been pointed out that in moderately asymmetrical (or skewed) distributions there tends

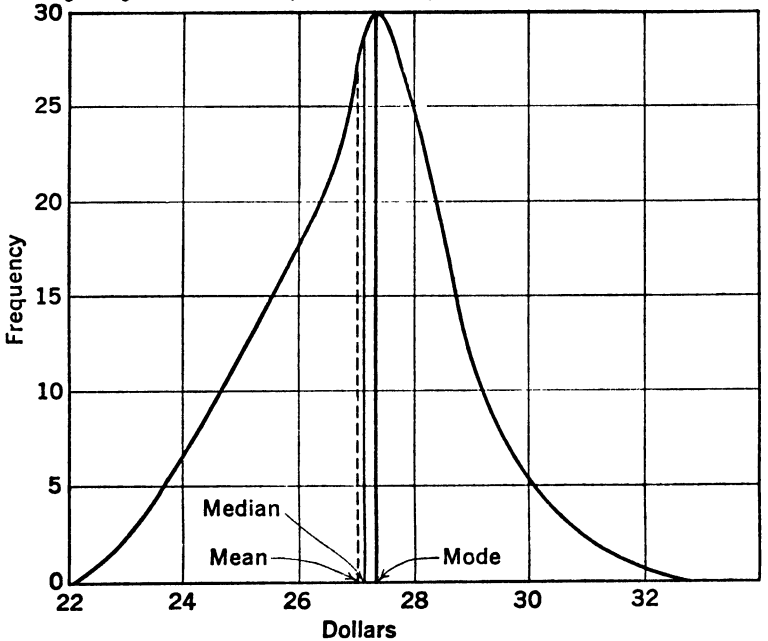


FIG. 47. — Distribution of Weekly Earnings of Employees. A Smoothed Frequency Curve, showing the Relation between Mean, Median and Mode

to be a constant relationship between the three averages which have been described, the median lying between the mean and the mode, and approximately one third of the distance from the former towards the latter. In the present case this relationship holds fairly well when the value of the mode is approximated from the smoothed curve. The irregularities in the original data render the process of smoothing by inspection rather arbitrary, however.

In Fig. 48 the same data are represented by a cumulative frequency curve, based upon Table 28 on page 124. The steepness of a cumulative frequency curve within any given inter-

val depends upon the number of cases added within the corresponding interval on the horizontal scale. Thus the curve rises gradually at first, then more steeply, and tails off gradually at the upper extremity. The value of the mode, obviously, is the reading on the horizontal scale corresponding to the point of greatest steepness. This is the point at

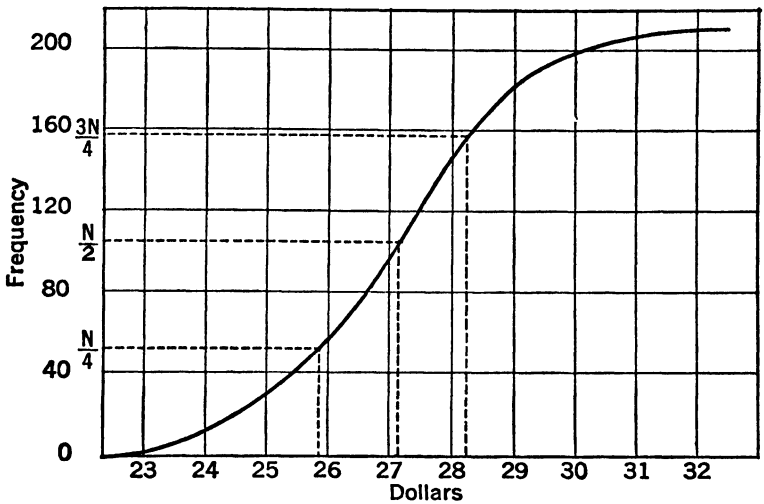


FIG. 48. — Cumulative Distribution of Weekly Earnings of Employees, Illustrating the Graphic Location of Median and Quartiles

which the increase of frequencies is greatest, the point of greatest concentration in the frequency distribution. The value of the mode may be approximated from a smoothed frequency curve by locating the point at which the slope is greatest (which is a point of inflection) and taking the corresponding reading on the x-scale. In the present case a value of approximately \$27.50 is secured for the mode by this method.

Values for the median, quartiles, and deciles may also be secured graphically from the cumulative frequency curve. The smoothing of such a curve provides a quite satisfactory method of interpolation and, if the scale of the diagram is sufficiently large, accurate values may be obtained by this method. Locate on the vertical scale (the scale of

TABLE 28

Cumulative Distribution of Wage-Earners in a Manufacturing Establishment

(Classified on the basis of weekly earnings)

Weekly earnings	*Number earning stated amount (frequency)*
Less than $22.50	0
" " 23.00	1
" " 23.50	5
" " 24 00	8
" " 24.50	19
" " 25.00	29
" " 25.50	41
" " 26.00	56
" " 26.50	78
" " 27.00	98
" " 27.50	122
" " 28.00	152
" " 28.50	169
" " 29 00	186
" " 29 50	193
" " 30 00	199
" " 30.50	204
" " 31.00	208
" " 31.50	209
" " 32.00	209
" " 32.50	210

cumulative frequencies) a point distant from the base by $\frac{N}{2}$.

If from this point a horizontal line be extended to the cumulative curve, the abscissa of the point of intersection will be the value of the median. This value may be easily determined by dropping a vertical line from the point of intersection to the x-axis. Figure 48 illustrates the application of this method. A value of $27.125 is secured for the median by this method. By direct interpolation a value of $27.1458 is obtained. The quartiles may be located in precisely the same way, the vertical scale being divided into quarters and horizontal lines extended to the cumulative curve from the points thus located on the vertical scale.

For some purposes, particularly those that involve the averaging of *rates* or *ratios* rather than quantities, none of the averages which have been described is suitable. The geometric and the harmonic means are types of averages that should be familiar because they are particularly appropriate for such purposes.

The Geometric Mean

The geometric mean is the nth root of the product of n measures; its value thus is represented by:

$$M_g = \sqrt[n]{a_1 \cdot a_2 \cdot a_3 \ldots a_n}.$$

The geometric mean of the numbers 2, 4, 8, is

$$\begin{aligned} M_g &= \sqrt[3]{2 \times 4 \times 8} \\ &= \sqrt[3]{64} \\ &= 4. \end{aligned}$$

It is obvious from the method of computation that if any one of the measures in the series has a value of zero the geometric mean is zero.

The actual computation of the geometric mean is greatly facilitated by the use of logarithms. In this form

$$\operatorname{Log} M_g = \frac{\log a_1 + \log a_2 + \log a_3 + \ldots + \log a_n}{N}.$$

The logarithm of the geometric mean is equal to the arithmetic mean of the logarithms of the individual measures.

When the measures, of which the geometric mean is desired, are to be weighted, the separate weights are introduced as exponents of the terms to which they apply. Thus if we represent the sum of the weights by N and the weights corresponding to the terms a_1, a_2, $a_3 \ldots a_n$, respectively, by w_1, w_2, $w_3 \ldots w_n$, the formula for the geometric mean is

$$M_g = \sqrt[n]{a_1^{w_1} \cdot a_2^{w_2} \cdot a_3^{w_3} \ldots a_n^{w_n}}.$$

This is equivalent to repeating each term a number of times, the number corresponding to the amount by which

it is weighted. (This, of course, is precisely what is done in securing a weighted arithmetic mean.) When logarithms are employed the formula for the weighted geometric mean becomes

$$\text{Log } M_g = \frac{w_1 \log a_1 + w_2 \log a_2 + w_3 \log a_3 + \ldots + w_n \log a_n}{N}.$$

A method of computing the geometric mean may be illustrated with reference to Table 29, which shows the distribution of the prices of 66 preferred stocks paying seven per cent dividends. The table is based upon closing prices on the New York Stock Exchange and the New York Curb Exchange for the week ended July 25, 1936.

TABLE 29

Computation of the Geometric Mean of Preferred Stock Prices

Class-interval	m	f	log m	f log m
\$ 70–\$ 89 9	80	5	1.90309	9.51545
90– 109.9	100	20	2 00000	40.00000
110– 129 9	120	27	2.07918	56.13786
130– 149 9	140	6	2.14613	12.87678
150– 169.9	160	8	2.20412	17.63296
		66		136.16305

$$\text{Log } M_g = \frac{136.16305}{66} \qquad \text{Log } M_g = 2.06308$$

$$M_g = 115.63$$

CHARACTERISTICS OF THE GEOMETRIC MEAN

The nature of the geometric mean may be understood by considering its relation to the terms it represents, as an average.

If the arithmetic mean of a series of measures replace each item in the series, the *sum* of the measures will remain unchanged. Thus, the sum of the numbers 2, 4, 8 is 14. The arithmetic mean of these three numbers is $4\frac{2}{3}$; if this value be inserted in the place of each of the three measures the sum remains 14. It is characteristic of the geometric

mean that the *product* of a series of measures will remain unchanged if the geometric mean of those measures replace each item in the series. Thus the product of 2, 4, 8 is 64. The geometric mean of the three numbers is 4; if this value replace each of the three measures the product remains 64.

Again, it is true of the arithmetic mean that the sum of the deviations of the items above the mean equals the sum of the deviations of the items below the mean (disregarding signs). The sums of the differences between the individual items and the mean are equal. In the case of the geometric mean the products of the corresponding ratios are equal. If the ratios of the geometric mean to the measures which it exceeds be multiplied together, the product will equal that secured by multiplying together the ratios to the geometric mean of the measures exceeding it in value. For example, the geometric mean of the numbers 3, 6, 8, 9 is 6. The following equation may be set up:

$$\frac{6}{3} \times \frac{6}{6} = \frac{8}{6} \times \frac{9}{6}$$

The last example brings out the most important characteristic of the geometric mean. It is a means of averaging ratios. Its chief use in the field of economic statistics has been in connection with index numbers of prices, where rates of change are of major importance. A rise in prices represented by the change from 50 to 100 is as important as a rise from 100 to 200. Yet this equivalence is not brought out by the arithmetic mean, which gives double weight to the change which involves an absolute difference of 100. An example frequently cited is that of two cases of price change, one a ten-fold increase, from 100 to 1,000, the other a fall to one tenth of the old price, from 100 to 10. The arithmetic mean of 1,000 and 10 is 505, the geometric mean is $\sqrt{1{,}000 \times 10}$, or 100. When the average is of the latter type it is seen that the two equal ratios of change have balanced each other. The arithmetic mean, 505, is

quite incorrect as a measure of average ratio of price change. This subject is discussed at greater length in the chapter on index numbers.

What has been said in an earlier section in regard to the advantages of logarithmic charting for certain purposes bears upon the use of the geometric mean. This average is sometimes called the logarithmic mean, as its logarithm is simply the arithmetic mean of the logarithms of the constituent measures. Wherever percentages of change are being averaged, where ratios rather than absolute differences are significant, the use of the geometric mean is advisable.

A problem involving the use of the geometric mean arises in computing the average rate of increase of any sum at compound interest. If p_o represent the principal at the beginning of the period, p_n the principal at the end of the period, r the rate of interest and n the number of years in the period, the sum to which p_o will amount at the end of the n years, if interest is compounded annually, is represented by the equation:

$$p_n = p_o(1 + r)^n.$$

It follows from this that:

$$r = \sqrt[n]{\frac{p_n}{p_o}} - 1.$$

Thus, if \$1,000 at compound interest amounts to \$1,600 at the end of 12 years, there has been an increase of 60 per cent. The arithmetic mean is 5 per cent, but this is not the rate at which the money increased. The true rate is:

$$\begin{aligned} r &= \sqrt[12]{\frac{1,600}{1,000}} - 1 \\ &= \sqrt[12]{1.60} - 1 \\ &= 1.04 - 1 \\ &= .04, \text{ or } 4\% \end{aligned}$$

Precisely the same problem arises whenever rates of increase or decrease are to be averaged. The use of the arithmetic mean gives an incorrect result.

THE GEOMETRIC MEAN AS A MEASURE OF CENTRAL TENDENCY

A question arises as to the type of frequency distribution the central tendency of which would be best represented by the geometric mean. When the absolute measures, plotted on the arithmetic scale, give a fairly symmetrical distribution, the arithmetic mean is clearly preferable to the geometric mean. But when the absolute figures thus plotted give an asymmetrical frequency curve of such a type that the asymmetry would be removed and a symmetrical curve secured by plotting the logarithms of the measures, the geometric mean would appear to be preferable. Such a distribution would be one in which not the absolute deviations about the central tendency but the *relative* deviations, the deviations as ratios, were symmetrical. The arithmetic mean of the logarithms of the various measures (which value is, as has been shown, the logarithm of the geometric mean of the original measures) would be the best representative of the central tendency in such a distribution. The curve thus plotted would be symmetrical about the logarithm of the geometric mean. A frequency curve representing the logarithms of percentage changes in prices would tend to show this symmetry about the logarithm of the geometric mean of these changes. These percentage changes, as natural numbers, group themselves in an asymmetrical form, with the range of deviations above the arithmetic mean greatly exceeding the range below.[1] This arises, of course, from the fact that prices of given commodities may increase 1,000 per cent or more from a given base, but cannot fall more than 100 per cent from any given base. The section

[1] Cf. Fig. 51.

on index numbers contains a fuller discussion of this particular phase of the subject.[1]

The construction of a frequency distribution in which logarithms are tabulated would be laborious, if the logarithm of each item to be entered had to be determined, before tabulation. It is possible, however, with no great trouble to construct a true logarithmic distribution, with class-interval constant in terms of logarithms. The 66 quotations on preferred stocks, tabulated in Table 29, range from 74 to 166. The logarithm of 74 is 1.86923; the logarithm of 166 is 2.22011. The range, in logarithms, is .35088. We may select .06 as a suitable logarithmic class-interval, for the present purpose. For convenience in tabulating the data we set up two series of class limits, one in terms of logarithms, one in terms of the corresponding natural numbers. In constructing the distribution natural numbers may be tabulated, utilizing the class limits defined in natural terms. All subsequent calculations may be carried through in terms of logarithms. The distribution appears in Table 30 on page 131.

If the geometric mean is considered appropriate for a given series, the type of distribution represented by Table 30 is more logical than that shown in Table 29, and the descriptive measurements secured from Table 30 have correspondingly greater validity. We may derive the mean of the

[1] C. M. Walsh, in *The Problem of Estimation* (London, P. S. King & Son, 1921) 35, lays down the following criteria for the use of averages:

(*a*) When there are no conceivable or assignable upper or lower limits to the values of the terms in a series, the arithmetic average should be employed.

(*b*) When there is a definite lower limit at or above zero and no upper conceivable or assignable limit, the geometric average should be employed. Because this is true of price changes Walsh believes the geometric average to be the correct one to use in making index numbers of prices.

(*c*) When in practice, or in the nature of things, certain upper and lower limits are found to exist and the above criteria cannot be employed, a study of the actual dispersion of the data is necessary. In this case, if the mode is found nearer to the arithmetic average, that average should be employed; if the mode is found nearer to the geometric average, that average should be used.

TABLE 30

Distribution of Prices of Preferred Stocks Paying Seven Per Cent Dividends

Class-interval (natural numbers)	*Class-interval (logarithms)*	*Mid-point (logarithms)*	*Frequency*	
		m	*f*	*fm*
$ 70.80–$ 81.27	1.85–1.9099	1.88	2	3.76
81.28– 93.32	1.91–1.9699	1.94	4	7.76
93.33– 107.15	1.97–2.0299	2 00	12	24.00
107.16– 123 02	2.03–2.0899	2.06	30	61.80
123.03– 141 24	2 09–2.1499	2.12	6	12.72
141.25– 162.17	2.15–2 2099	2.18	7	15 26
162.18– 186.20	2.21–2.2699	2.24	5	11.20
			66	136.50

logarithms of the preferred stock prices by dividing Σfm of Table 30 (136.50) by 66. The value is 2.06818. The anti-log of this is 116.97, which is the geometric mean of the distribution. This differs somewhat from the value $115.63 secured from Table 29. The difference is due, in part, to the use of different class-intervals and class limits in the two cases. With a relatively small number of observations such differences would be expected to lead to different results. Differing assumptions concerning the internal distribution of items within the several classes would also contribute to a discrepancy between the two results. The value obtained from Table 30 is probably a closer approximation to the actual geometric mean than that obtained from Table 29.

A frequency curve based upon the logarithms of the measures included rather than upon the natural numbers, has been employed to advantage in plotting data relating to income distribution. When natural numbers are plotted, the range of income distribution is so large that it is physically impossible to prepare a chart that will reveal the characteristic features of all sections of the curve. The process of plotting on double logarithmic paper (which is, of course, equivalent to plotting the logarithms of both x's and y's)

meets this difficulty, giving a true impression of the whole distribution and the relations between its parts, and, at the same time, brings out certain important features that are obscured in the natural scale chart. In particular, this device appears to smooth into a straight line that part of the curve lying above the mode, a fact which led Vilfredo Pareto to enunciate what has been known as Pareto's Law concerning income distribution. An intensive study of the distribution of income in the United States has led the staff of the National Bureau of Economic Research to call into question certain conclusions drawn from Pareto's generalizations, though the value of the double logarithmic scale for the presentation of income data has been recognized.

The Harmonic Mean

The harmonic mean is a type of average capable of application only within a restricted field, but which should be employed to avoid error in handling certain types of data. It must be used in the averaging of time rates and it has distinctive advantages in the manipulation of some types of price data. The following example will illustrate the method of employing this average.

A given commodity is priced, in three different stores, at "four for a dollar," "five for a dollar" and "twenty for a dollar." The average price per unit is required. The arithmetic average of the figures given (4, 5, and 20) is $9\frac{2}{3}$. If we take this to be the average number sold per dollar, the average price would appear to be \$1.00 $\div$ $9\frac{2}{3}$, or $10\frac{10}{29}$ cents each. But the original quotations are equivalent to unit prices of 25 cents, 20 cents, and 5 cents; the arithmetic average of these prices is $16\frac{2}{3}$ cents apiece. The discrepancy between $10\frac{10}{29}$ cents and $16\frac{2}{3}$ cents is due to a faulty use of the arithmetic mean in averaging quotations in the "so many per dollar" form. Such a mean is, in effect, a weighted average, with greater weight being given to quotations involving a larger number of commodity units.

The correct result may be secured by taking the harmonic mean of the three original quotations. *The harmonic mean of a series of numbers is the reciprocal of the arithmetic mean of the reciprocals of the individual numbers.* Thus if we represent the numbers to be averaged by $r_1, r_2 \ldots r_n$, the formula for the harmonic mean, H, is

$$\frac{1}{H} = \frac{\frac{1}{r_1} + \frac{1}{r_2} + \frac{1}{r_3} + \ldots + \frac{1}{r_n}}{N}.$$

Using the figures just quoted:

$$\frac{1}{H} = \frac{\frac{1}{4} + \frac{1}{5} + \frac{1}{20}}{3}$$

$$= \frac{10}{60} = \frac{1}{6}$$

$$H = 6.$$

The harmonic mean of 4, 5, and 20 is 6, the average number of units sold per dollar. The average price per unit is $16\frac{2}{3}$ cents.

The computation of the harmonic mean of a series of magnitudes is greatly facilitated by the use of prepared tables of reciprocals.[1]

Relations between Different Averages

When different averages are located or computed for a given series of magnitudes, certain relationships between them are found to prevail.

1. The arithmetic mean, the median and the mode coincide in a symmetrical distribution.
2. In a moderately asymmetrical distribution the median lies between the mean and the mode, approximately one third of the distance along the scale from the former towards the

[1] *Barlow's Tables of Squares, Cubes, Square Roots, Cube Roots and Reciprocals*, New York, Spar and Chamberlain.

latter. Hence, for this type of distribution there is an approximation to the following relationship:

$$Mo = M - 3(M - Md).$$

3. The arithmetic mean of any series of magnitudes is greater than their geometric mean.
4. The geometric mean of any series of magnitudes is greater than their harmonic mean. The only exception to the last two rules is found when all the measures in the series are equal, in which case arithmetic mean, geometric mean and harmonic mean are equal.
5. The geometric mean of any two terms is equal to the geometric mean of the harmonic and arithmetic means of those terms. Thus if the terms be 2 and 8, the harmonic mean is $3\frac{1}{5}$, the geometric mean 4, and the arithmetic mean 5. But 4 is also the geometric mean of $3\frac{1}{5}$ and 5. This relationship does not hold when the series includes more than two terms, unless the terms constitute a geometric series.
6. When the dispersion of data follows the arithmetic law, the mode and median will generally be found closer to the arithmetic than to the geometric average. When the dispersion follows the geometric law the mode and median will generally be found closer to the geometric than to the arithmetic average.

Characteristic Features of the Chief Averages

The arithmetic mean

1. The value of the arithmetic mean is affected by every measure in the series. For certain purposes it is too much affected by extreme deviations from the average.
2. The arithmetic mean is easily calculated, and is determinate in every case.
3. The arithmetic mean is a computed average, and hence is capable of algebraic manipulation.

The median

1. The value of the median is not affected by the magnitude of extreme deviations from the average.
2. The median may be located when the items in a series are not capable of quantitative measurement.
3. The median may be located when the data are incomplete, provided that the number and general location of all the

cases be known, and that accurate information be available concerning the measures near the center of the distribution.
4. The median is not as well adapted to algebraic manipulation as the arithmetic, geometric and harmonic means.

The mode

1. The value of the mode is not affected by the magnitude of extreme deviations from the average.
2. The approximate mode is easy to locate but the determination of the true mode requires extended calculation.
3. The mode has no significance unless the distribution includes a large number of measures and possesses a distinct central tendency.
4. The mode is the average most typical of the distribution, being located at the point of greatest concentration.
5. The mode is not capable of algebraic manipulation.

The geometric mean

1. The geometric mean gives less weight to extreme deviations than does the arithmetic mean.
2. It is strictly determinate in averaging positive values.
3. The geometric mean is the form of average to be used when rates of change or ratios between measures are to be averaged, as equal weight is given to equal ratios of change. It is particularly well adapted to the averaging of ratios of price change.
4. The geometric mean is capable of algebraic manipulation.

The harmonic mean

1. The harmonic mean is adapted to the averaging of time rates and certain similar terms. It has been employed in the field of economic statistics in the manipulation of price data.
2. The labor of computing the harmonic mean and its unfamiliarity detract from its usefulness in ordinary statistical analysis.
3. The harmonic mean is capable of algebraic manipulation.

This summary has been designed to show that each type of average has its own particular field of usefulness. Each one is best for certain purposes and under certain conditions. The characteristics and limitations of each one should be understood in order that it may be appropriately employed. A complete description of a frequency distribu-

tion frequently calls for the determination of two or three of the chief averages, as well as other statistical measurements. The arithmetic mean is perhaps the most useful single average. The simplicity of its computation, the possibility of employing it in algebraic calculations and the fact that its meaning is perfectly definite and familiar make it highly serviceable in statistical work. Its sphere of usefulness is not universal, however, and it should only be employed when the given conditions render it suitable. A fuller appreciation of the distinctive virtues of the geometric mean is leading to a wider employment of that measure in many types of statistical work. A discriminating use of averages is essential to sound statistical analysis.

REFERENCES

Bowley, A. L., *Elements of Statistics*, Chap. 5.

Chaddock, R. E., *Principles and Methods of Statistics*, Chaps. 6–8.

Croxton, F. E. and Cowden, D. J., *Practical Business Statistics*, Chaps. 9, 10.

Davies, George R. and Crowder, W. F., *Methods of Statistical Analysis in the Social Sciences*, Chap. 3.

Davies, George R. and Yoder, D., *Business Statistics*, Chap. 2.

Day, Edmund E., *Statistical Analysis*, Chaps. 9, 10.

Davis, Harold T. and Nelson, W. F. C., *Elements of Statistics*, Chap. 3.

Jones, D. C., *A First Course in Statistics*, Chaps. 4, 5.

Kelley, Truman L., *Statistical Method*, Chap. 3.

Richardson, C. H., *An Introduction to Statistical Analysis*, Chap. 3.

Riegel, Robert, *Elements of Business Statistics*, Chaps. 10, 11.

Riggleman, John R. and Frisbee, Ira N., *Business Statistics*, Chap. 7.

Secrist, Horace, *Introduction to Statistical Methods*, Chap. 9.

Walsh, C. M., *The Problem of Estimation.*

Waugh, A. E., *Elements of Statistical Method*, Chap. 4.

Yule, G. U. and Kendall, M. G., *An Introduction to the Theory of Statistics*, Chap. 7.

Zizek, Franz, *Statistical Averages.*

CHAPTER V

DESCRIPTION OF THE FREQUENCY DISTRIBUTION: MEASURES OF VARIATION AND SKEWNESS

In the preceding chapters we have been concerned, first, with methods of reducing a mass of quantitative data to a form in which the characteristics of the mass as a whole may be readily determined and, in the second place, with methods of describing the assembled data. The first object is accomplished with the formation of a frequency distribution. The second is partially accomplished when there has been obtained a single significant value in the form of an average which represents the central tendency of the distribution. But any average, by itself, fails to give a complete description of a frequency distribution. Three other values are needed before the chief characteristics of a given distribution have been measured, and comparison with other distributions is possible. The first of these is a measure of the degree to which the items included in the original distribution depart or *vary* from the central value, the degree of "*scatter*," *variation* or *dispersion*. The second is a measure of the degree of symmetry of the distribution, of the balance or lack of balance on the two sides of the central value. The third is a measure of *kurtosis*, of the degree to which there is a bunching of cases at the modal value. The present chapter deals with various measures of variation and skewness. The method of measuring kurtosis is referred to at a later point.

Nature and Significance of Variation

The fact of variation in collections of quantitative data has been pointed out in earlier sections and the bearing of

this fact upon the work of the statistician indicated. Practically every collection of quantitative data, consisting of measurements from the social, biological, or economic field, is characterized by variation, by quantitative differences among the individual units. And this fact of variation is as important as the fact of family resemblance. Biological variation has been a fundamental factor in the evolutionary process. No measurement of a physical characteristic of a racial group, such as height, is complete without an accompanying measure of the average variation in the group in this respect. The average income in a country is perhaps of less significance than the variation in income, the differences between the incomes received by different economic classes. Price variations interrupt the normal functioning of the economic system, causing hardship to some and giving unearned profits to others, because the various elements in the price system are unequally affected. Not changes in the general level of prices but differences among changes in the prices of individual commodities and services cause trouble.

An average, by itself, has little significance unless the degree of variation in the given frequency distribution is known. If the variation is so great that there is no pronounced central tendency an average has no significance. With a decrease in the degree of variation an average becomes increasingly significant. Whether a single frequency distribution is being described, therefore, or comparison is being made with other distributions, a measure of central tendency must be supplemented by a measure of variation.

Measures of Absolute Variation

Variation may be expressed in terms of the units of measurement employed for the original data, or may be expressed as an abstract figure, such as a percentage, which is independent of the original units. When the original

units are employed *absolute variability* is measured; when an abstract figure is secured we have a measure of *relative variability*, more suitable for comparison than the former type. Measures of absolute variability are first considered.

THE RANGE

A rough measure of variation is afforded by the *range*, which is the absolute difference between the value of the smallest item and the value of the greatest item included in the distribution. Table 20 in Chapter IV shows the distribution of London-New York monthly exchange rates during the period 1882–1913. The smallest item among the original figures included in the table is \$4.83; the greatest is \$4.908. The range, therefore, is \$4.908–\$4.83, or \$.078. A distance on the scale equal to \$.078 will include every item. If the original data were not to be had the range could be approximated from the frequency table. It would be the difference between the lower limit of the class at the lower extreme of the distribution, and the upper limit of the class at the upper extreme, or \$.085 in the present case.

The value of the range, it is obvious, depends upon the values of the two extreme cases only. A single abnormal item would change its value materially. Because it is erratic and is likely to be unrepresentative of the true distribution of items, it is seldom used in statistical work. The range is frequently employed as a measure of stock market fluctuations, though its adequacy for this purpose may be questioned.

THE MEAN DEVIATION

A more accurate measure of the dispersion of items about a central value is afforded by the simple device of measuring the deviation of each item from this central value and averaging these deviations. The simple example in Table 31 illustrates the method of computation:

TABLE 31

Computation of Mean Deviation

x	f	d	
3	1	6	$M = 9$
6	1	3	
9	1	0	$M.D. = \frac{18}{5} = 3.6$
12	1	3	
15	1	6	
		18	

The average (the mean and median coincide in this case) is 9. The deviations are added, taking no account of algebraic signs, and the total divided by the number of items. This procedure is described by the expression

$$M.D. = \frac{\Sigma d}{N}.$$

In general terms, the *mean deviation* of a series of magnitudes is the arithmetic mean of their deviations from an average value (either mean or median). In the process of summation and averaging the algebraic signs of the deviations are disregarded. In practice it makes little difference whether deviations be measured from the mean or the median. Theoretically the latter should be chosen, for the value of the mean deviation is least when the median is the point of reference.

Table 32 illustrates the computation of the mean deviation when the data are grouped in a frequency distribution.[1] In this work, as in certain other computations, we make the assumption that the items in each class-interval are uniformly distributed throughout that interval.

The median hourly wage of the 4,216 steel workers represented in this distribution is 48.11 cents. The mean devia-

[1] Since the uses of the mean deviation are somewhat limited, the beginning student may well omit the remainder of the section on the mean deviation. After study of the more widely employed standard deviation the student may wish to return to the computation of the mean deviation of observations grouped in a frequency distribution.

TABLE 32. *Computation of Mean Deviation*
Average Hourly Earnings of Workers in Open-Hearth Furnaces in the Great Lakes and Middle West District in 1933

Class-interval (in cents per hour)	*Mid-point*	*Frequency*	*Deviation from arbitrary origin*		
	m	f	d'	fd'	
25.0– 29.9	27.5	12	20	240	$c = 0.61$
30.0– 34.9	32.5	472	15	7,080	(Median = 48.11
35.0– 39.9	37.5	700	10	7,000	Arbitrary origin = 47.5
40.0– 44.9	42.5	601	5	3,005	$c = 48.11 - 47.5 = 0.61$)
45.0– 49.9	47.5	520	0	0	
50.0– 54.9	52.5	537	5	2,685	N_a = No. of observations in
55.0– 59.9	57.5	397	10	3,970	classes above that
60.0– 64.9	62.5	225	15	3,375	containing the median
65.0– 69.9	67.5	139	20	2,780	= 1911
70.0– 74.9	72.5	111	25	2,775	
75.0– 79.9	77.5	43	30	1,290	N_b = No. of observations in
80.0– 84.9	82.5	111	35	3,885	classes below that
85.0– 89.9	87.5	74	40	2,960	containing the median
90.0– 94.9	92.5	59	45	2,655	= 1785
95.0– 99.9	97.5	45	50	2,250	
100.0–104.9	102.5	51	55	2,805	N_m = No. of observations in
105.0–109.9	107.5	78	60	4,680	the class-interval con-
110.0–114.9	112.5	6	65	390	taining the median
115.0–119.9	117.5	17	70	1,190	= 520
120.0–124.9	122.5	1	75	75	
125.0–129.9	127.5	2	80	160	$i = 5$
130.0–134.9	132.5	5	85	425	
135.0–139.9	137.5	7	90	630	*Calculations*
140.0–144.9	142.5	1	95	95	
145.0–149.9	147.5	1	100	100	(1) Sum of deviations from
150.0–154.9	152.5	0	105	0	arbitrary origin of all
155.0–159.9	157.5	1	110	110	observations in classes
		4,216		56,610	other than that contain-
					ing the median = 56,610

Computation of median:

$$\frac{N}{2} = 2{,}108$$

$$Md = 45.0 + \left\{\frac{323}{520} \times (5.0)\right\}$$

$$= 45.0 + 3.11$$

$$= 48.11$$

(2) $(N_b - N_a)c = -76.86$

(3) $N_m \dfrac{\left(\frac{i}{2} + c\right)^2}{2i}$

$$+ N_m \frac{\left(\frac{i}{2} - c\right)^2}{2i} = 688.67$$

Sum of deviations from median = 56,610 − 76.86
+ 688.67
= 57,221.81

$$M.D. = \frac{57{,}221.81}{4{,}216}$$

$$= 13.573$$

tion could be computed directly, with reference to deviations from the median, but it is simpler to measure the deviations from the midpoint of the class containing the median, and then apply corrections to offset the resulting error.

In Table 32 deviations have been measured not from 48.11, the value of the median, but from 47.5, the midpoint of the class in which the median falls. Working with these measurements, the computations involve three steps:

1. Obtaining the sum of the deviations from the assumed median of all items falling in classes other than that containing the true median.

2. Correcting this sum for the error involved in the use of an origin other than the true median.

3. Adding to the corrected sum the sum of the deviations from the median of the items within the class-interval containing the median.

(1) The sum referred to in (1) is obtained directly, in the manner indicated in Table 32.[1] It comes to 56,610.

(2) The four classes below that containing the median contain 1,785 items. The deviation of each of these items from the true median, 48.11, is greater by 0.61 than the deviations actually recorded in Table 32, which are measured from 47.5. The measured deviations are too small by 0.61 for 1,785 items. The 22 classes above that containing the median contain 1,911 items. For each of these the deviation from the true mean, 48.1, is less by 0.61 than the deviations actually recorded, which are measured from 47.5. The measured deviations are too large by 0.61 for 1,911 items. Accordingly the figure 56,610, which we have obtained as the sum of the deviations from the arbitrary origin of all the items in classes other than that containing the true median, must be corrected by the addition to it, algebraically, of $+ (1{,}785 \times 0.61)$ and $- (1{,}911 \times 0.61)$.

[1] This is not the sum of the deviations from 4.75, the arbitrary origin. For no account is taken of the deviations from that value of the 520 items falling within the class in question. If these are scattered uniformly throughout the class-interval they will contribute to the total of the deviations from 4.75. This would not be so if we were working on the assumption that all the items in a class are concentrated at the midpoint. In computing the mean deviation, however, it is necessary to make a different assumption, namely, that of uniform distribution throughout the class-interval.

The corrections under point (2) may be defined more briefly.[1] Let N_a = number of items in classes above that containing the median, N_b = number of items in classes below that containing the median, and $c = Md - O$, where Md is the median and O is the arbitrary origin. The quantity c will, of course, be positive or negative, depending on the relative values of Md and O, and this sign should be retained throughout the calculations. The correction noted in (2) is then given by

$$(N_b - N_a)c$$

which is to be added (algebraically) to the sum referred to in (1). In the present instance we have, as the required correction,

$$(1{,}785 - 1{,}911) \times (+0.61) = -76.86.$$

(3) Taking account of point (3) now, we must measure the deviations from the median of the 520 observations hitherto neglected. These are the observations falling within the class-interval that contains the median. This class-interval extends, on the x-scale, from 45.0 to 50.0. The value of the median is 48.11. If the 520 observations are uniformly distributed between 45.0 and 50.0, the number falling between 45.0 and 48.11 may be computed by the direct proportion

$$\frac{3.11}{5.0} \times 520 = 323.4.$$

Similarly, for the number of observations between 48.11 and 50.0, we have

$$\frac{1.89}{5.0} \times 520 = 196.6.$$

On the assumption of uniform distribution, the average deviation from the median of the 323.4 observations falling between 45.0 and 48.11 is 1.555 $\left(\text{i.e., } \frac{3.11}{2}\right)$. For the sum of the deviations of this group from the median, we have

$$323.4 \times 1.555 = 502.887.$$

Similarly, the average deviation from the median of the 196.6 observations falling between 48.11 and 50.0 is .945 $\left(\text{i.e., } \frac{1.89}{2}\right)$.

[1] Cf. *A Handbook of Mathematical Statistics*, H. L. Rietz, editor, Boston, Houghton Mifflin, 1924, 30.

For the sum of the deviations of this group from the median we have

$$(196.6) \times .945 = 185.787.$$

The sum of the deviations from the median of all the observations in the class containing the median is

$$502.887 + 185.787 = 688.674.$$

In more general terms, the correction noted in (3) may be defined as follows. We have $c = Md - O$; let i = class-interval and let N_m = number of observations in the class-interval in which the median lies. The sign of c must be retained in the calculations. For the number of items in that portion of this class-interval which falls below the median, we have

$$\left(\frac{\frac{i}{2}+c}{i}\right)N_m.$$

The average deviation of these items from the median is

$$\frac{\frac{i}{2}+c}{2}.$$

The sum of the deviations from the median of the items in this segment of the class-interval containing the median is the product of these two quantities, or

$$N_m\left(\frac{\frac{i}{2}+c}{i}\right) \times \frac{\frac{i}{2}+c}{2} = N_m\frac{\left(\frac{i}{2}+c\right)^2}{2i}.$$

For the number of items in that portion of this class-interval which lies above the median, we have

$$\left(\frac{\frac{i}{2}-c}{i}\right)N_m.$$

The average deviation of these items from the median is

$$\frac{\frac{i}{2}-c}{2}.$$

The sum of the deviations from the median of the items in this segment of the class-interval containing the median is given by

$$N_m \frac{\left(\frac{i}{2} - c\right)^2}{2i}.$$

Accordingly, the total correction referred to under (3), on p. 142, or the sum of the deviations from the median of the items within the class-interval containing the median, is

$$N_m \frac{\left(\frac{i}{2} + c\right)^2}{2i} + N_m \frac{\left(\frac{i}{2} - c\right)^2}{2i}.$$

The nature of these formulas may be made clearer by insertion of the values in the example cited above.

In the final computation of the mean deviation we must apply to the sum referred to under (1), on p. 142, the two corrections noted under (2) and (3) on p. 142. From (1) we have 56,610; the correction under (2) is − 76.86; the correction under (3) is + 688.67. The sum of the deviations from the median is, therefore, 57,221.81. For the mean deviation from the median, we have

$$M.D. = \frac{57{,}221.81}{4{,}216} = 13.573.$$

The mean deviation from the mean may be computed by an identical process.

THE STANDARD DEVIATION

The process of calculating the mean deviation is algebraically illogical because algebraic signs are disregarded. In the computation of the standard deviation this error is avoided and a measure of more precise mathematical significance is secured. The conventional symbol for the standard deviation is the Greek letter *sigma*, σ.

In computing this measure the deviations of the individual items from the arithmetic mean are squared, totaled,

the mean of the squared deviations obtained, and the square root of this mean extracted. The *standard deviation* is, thus, the square root of the mean of the squared deviations. This measure is also termed the *root-mean-square deviation*, a useful name because it describes in full the method of calculation. The deviations are always measured from the arithmetic mean, as the value of the measure is a minimum under these conditions. A simple example will illustrate the process (Table 33).

TABLE 33

Computation of Standard Deviation

m	f	d	d^2
3	1	− 6	36
6	1	− 3	9
9	1	0	0
12	1	+ 3	9
15	1	+ 6	36
	5		90

$$M = 9$$

$$\sigma = \sqrt{\frac{90}{5}}$$

$$= \sqrt{18}$$

$$\sigma = 4.24$$

When the standard deviation is computed from ungrouped data, as in this case, the formula [1] is

$$\sigma = \sqrt{\frac{\Sigma d^2}{N}}.$$

When the items are grouped in a frequency distribution the task of computation is a little more complicated. The measurement of deviations from an arbitrary origin is essential in this case, as it greatly simplifies the calculations.

[1] This formula is used in statistical *description*, which is the concern of this section of the book. If our purpose is to use results secured from a sample as *estimates* of the attributes of the population from which the sample has been drawn, a slight modification is desirable. It has been shown that the estimate of the true standard deviation is improved if $N-1$ be used as the divisor in the formula, in place of N. The difference is slight for estimates based on large samples, important for small ones.

The general formula for the standard deviation is

$$\sigma = \sqrt{\frac{\Sigma f d^2}{N}}$$

where f represents the class-frequencies, d the deviations from the arithmetic mean and N the number of cases included. It follows that

$$\sigma^2 = \frac{\Sigma f d^2}{N}.$$

If a deviation from an arbitrary origin be represented by d' and the root-mean-square deviation from this origin be represented by s_a, we have

$$s_a{}^2 = \frac{\Sigma f (d')^2}{N}.$$

The root-mean-square deviation from the mean (σ) is less than the root-mean-square deviation from any other point on the scale. Hence $s_a{}^2$ is greater than σ^2. We may represent by c the difference between the true mean and the arbitrary origin. It may be readily established [1] that

$$\sigma^2 = s_a{}^2 - c^2.$$

The value of the standard deviation may be most easily determined, therefore, by computing $s_a{}^2$ and c^2. The operations involved are illustrated in detail in Table 34, showing the distribution of 11,404 steel workers, classified on the basis of average hourly earnings in 1933.

[1] For $\sigma^2 = \frac{\Sigma d^2}{N}$ but $\Sigma d = 0$

$s_a{}^2 = \frac{\Sigma (d')^2}{N}$ $\quad \therefore \Sigma (d')^2 = \Sigma d^2 + Nc^2$

$d' = d + c$ $\quad \frac{\Sigma (d')^2}{N} = \frac{\Sigma d^2}{N} + c^2$

$(d')^2 = d^2 + 2cd + c^2$ $\quad s_a{}^2 = \sigma^2 + c^2$

$\Sigma (d')^2 = \Sigma d^2 + 2c\Sigma d + Nc^2$ $\quad \sigma^2 = s_a{}^2 - c^2.$

TABLE 34. *Computation of Standard Deviation*
Average Hourly Earnings of Workers in Open-Hearth Furnaces in 1933

(1) Class-interval (cents per hour)	(2) Mid-point (cents)	(3) Frequency	(4) Deviation from arbitrary origin	(5)	(6)	(7)	(8)
	m	f	d'	fd'	$f(d')^2$	$(d'+1)^2$	$f(d'+1)^2$
15.0– 19.9	17.5	41	− 9	− 369	3,321	64	2,624
20.0– 24.9	22.5	54	− 8	− 432	3,456	49	2,646
25.0– 29.9	27.5	342	− 7	− 2,394	16,758	36	12,312
30.0– 34.9	32.5	1,158	− 6	− 6,948	41,688	25	28,950
35.0– 39.9	37.5	2,103	− 5	− 10,515	52,575	16	33,648
40.0– 44.9	42.5	2,063	− 4	− 8,252	33,008	9	18,567
45.0– 49.9	47.5	1,433	− 3	− 4,299	12,897	4	5,732
50.0– 54.9	52.5	1,131	− 2	− 2,262	4,524	1	1,131
55.0– 59.9	57.5	775	− 1	− 775	775	0	0
60.0– 64.9	62.5	478	0	0	0	1	478
65.0– 69.9	67.5	457	1	457	457	4	1,828
70.0– 74.9	72.5	304	2	608	1,216	9	2,736
75.0– 79.9	77.5	216	3	648	1,944	16	3,456
80.0– 84.9	82.5	193	4	772	3,088	25	4,825
85.0– 89.9	87.5	117	5	585	2,925	36	4,212
90.0– 94.9	92.5	111	6	666	3,996	49	5,439
95.0– 99.9	97.5	62	7	434	3,038	64	3,968
100.0–104.9	102.5	71	8	568	4,544	81	5,751
105.0–109.9	107.5	103	9	927	8,343	100	10,300
110.0–114.9	112.5	34	10	340	3,400	121	4,114
115.0–119.9	117.5	58	11	638	7,018	144	8,352
120.0–124.9	122.5	27	12	324	3,888	169	4,563
125.0–129.9	127.5	19	13	247	3,211	196	3,724
130.0–134.9	132.5	19	14	266	3,724	225	4,275
135.0–139.9	137.5	14	15	210	3,150	256	3,584
140.0–144.9	142.5	12	16	192	3,072	289	3,468
145.0–149.9	147.5	2	17	34	578	324	648
150.0–154.9	152.5	4	18	72	1,296	361	1,444
155.0–159.9	157.5	2	19	38	722	400	800
160.0–164.9	162.5	1	20	20	400	441	441
Total		11,404		− 28,200	229,012		184,016

$N = 11{,}404$

Class-interval = 5.0 cents

c (in class-interval units) $= \dfrac{-28{,}200}{11{,}404} = -2.4728$

c^2 (in class-interval units) $= +6.1147$

s_a^2 (in class-interval units) $= \dfrac{\Sigma f(d')^2}{N} = \dfrac{229{,}012}{11{,}404} = 20.0817$

σ^2 (in class-interval units) $= s_a^2 - c^2 = 20.0817 - 6.1147 = 13.9670$

σ (in class-interval units) $= 3.737$

σ (in original units) $= 3.737 \times 5.0$ cents $= 18.685$ cents.

The entire calculation, it will be noted, is carried through in terms of class-interval units, the result being reduced to the original units in the final operation. In computing c, the difference between the true mean and the arbitrary origin, the algebraic sum of the deviations is divided by the number of cases. The arithmetic mean could be determined by reducing c to original units and adding this value (algebraically) to the value of the arbitrary quantity selected as origin, but this is not an essential step. The actual value of the mean need not be known in the computation of the standard deviation.

A check upon the accuracy of the calculations (the *Charlier check* [1]) is afforded by the figures in cols. (7) and (8) of Table 34. If deviations be measured, not from the arbitrary origin employed in computing the standard deviation, but from an origin one class-interval below, we secure a set of values equal to $d' + 1$. The squares of these values are given in col. (7). Multiplying by the corresponding frequencies we have the quantities recorded in col. (8), the sum of which is 184,016. This total stands in a definite relationship to the values secured in computing the standard deviation. For

$$\Sigma f(d' + 1)^2 = \Sigma f[(d')^2 + 2d' + 1]$$
$$= \Sigma f(d')^2 + 2\Sigma fd' + \Sigma f$$

or

$$\Sigma f(d' + 1)^2 = \Sigma f(d')^2 + 2\Sigma fd' + N.$$

Inserting in this last equation the values secured from the calculations shown in Table 34, we obtain this check:

$$184{,}016 = 229{,}012 + 2(-28{,}200) + 11{,}404$$
$$= 184{,}016.$$

The following is a summary of the steps in the process of computing the standard deviation of items grouped in a frequency distribution:

[1] Cf. C. V. L. Charlier, *Vorlesungen Über Die Grundzüge Der Mathematischen Statistik*, Lund, Verlag Scientia, 1920, 19.

1. Select as arbitrary origin the mid-point of a class near the center of the distribution.
2. Measure the deviations from this point of the items in each class, in class-interval units. Multiply the deviations by the corresponding class-frequencies.
3. Divide the algebraic sum of the deviations by N. This gives c, in class-interval units. Compute c^2.
4. Square the deviations and multiply by the corresponding class-frequencies.
5. Divide the sum of the squared deviations by N. This gives s_a^2, in class-interval units.
6. From the formula, $\sigma^2 = s_a^2 - c^2$, compute σ^2. Extract the square root of this value, securing σ in class-interval units.
7. Multiply σ, as thus computed, by the class-interval. The result is σ in the original units of measurement.

Certain of the characteristics of the standard deviation and its relation to other measures of dispersion are described in a later section [1]

THE QUARTILE DEVIATION

In the chapter on averages methods of locating the quartiles and deciles were described. The former are those points on the scale of values, along which the items of a given distribution lie, which divide the total number of items into four equal groups. The deciles are those points dividing the total number of items into ten equal groups. The degree and character of the variation in a frequency distribution may be accurately described if the location of the quartiles and deciles is shown. Such knowledge, however, while helpful in giving a picture of the distribution, is not as useful for purposes of concise description and comparison as knowledge of the values of the mean deviation or the standard deviation. The significance of a single measure is more readily grasped than is the meaning of a number of interrelated values. Such a measure of variation may be computed from the quartiles, however. With regard to ease of

[1] A correction to be applied to the standard deviation in certain cases (Sheppard's correction) is described in Chapter XIII.

calculation and immediate significance this *quartile deviation* has distinct merits.

Within the range between the two quartiles, of course, one half of all the measures are included. The greater the concentration the smaller this interval, hence a fairly accurate measure of dispersion may be obtained from the relationship between these two quartiles. The quartile deviation is the *semi-interquartile range*, half the distance along the scale between the first and third quartiles. Thus if $Q.D.$ represent the quartile deviation, Q_1 the first quartile and Q_3 the third quartile,

$$Q.D. = \frac{Q_3 - Q_1}{2}.$$

If the value of a point on the scale half-way between the first and third quartiles is represented by K, one half of all the measures in a frequency distribution will fall within the range $K \pm Q.D.$ For the data in Table 32, relating to the hourly earnings in 1933 of steel workers in the Great Lakes and Middle West District, we have

$$\begin{aligned} Q_1 &= 39.07 \\ Q_3 &= 59.03 \\ Q.D. &= \frac{59.03 - 39.07}{2} \\ &= 9.98 \\ K &= 39.07 + 9.98 \\ &= 49.05. \end{aligned}$$

Thus one half of all the measures lie within the range 49.05 ± 9.98. This statement, together with a statement of the average hourly earnings in 1933 (mean, median, or mode), constitutes a useful description of the distribution. In a perfectly symmetrical distribution the value of K will coincide with the value of the median (that is, the median will lie half-way along the scale from Q_1 to Q_3). The distribution of wage rates is slightly asymmetrical, the value of

the median being 48.11, as compared with the value of 49.05 for K.

THE PROBABLE ERROR

In studying the results of astronomical and other physical measurements it has been found that the values secured by different observers for the same constant quantity vary. These varying results, however, are distributed in a certain definite way, and when plotted give a curve similar to the normal curve of error. In such cases there is an immediate and obvious need of some measure of variation which may be used as an index of the reliability of given results. If the results secured by different investigators, or by the same investigator at different times, vary widely they cannot be accepted as reliable, while the reverse is true if the variation is slight. The measure of dispersion which has been generally employed in such cases is termed the *probable error*. The probable error is that amount which, in a given case, is exceeded by the errors of one half the observations. Since the most probable value of a given series of observations is their arithmetic mean, the probable error is always measured from the mean. The name of this measure derives from the fact that the probability that a given observation will vary from the mean of all the observations by an amount greater than the probable error is exactly $\frac{1}{2}$. It follows that, when the observations are arranged in the form of a frequency distribution, a distance equal to the probable error laid off on each side of the arithmetic mean will define limits within which one half of the total number of cases will fall.

This measure of variation has been employed in fields other than that in which it was originally applied, fields in which the name *probable error* is somewhat misleading. In such cases it is perhaps better to think of it as the *probable deviation*, that distance from the mean which will be exceeded by one half of the total deviations.

The probable error is a measure of dispersion which is fully significant only when it applies to a distribution following the normal law of error. In such cases it has a definite and precise meaning. This is not so when it is applied to skew distributions, and its use in such cases is not advisable. The quartile deviation, the value of which is equal to that of the probable error in a normal distribution, has a more direct significance than the probable error in the description of abnormal distributions, and should be employed in such cases. In a later section the use of the probable error as a measure of the reliability of statistical results is more fully explained.

The value of the probable error in a given case, assuming a normal distribution to prevail, may be determined from the value of the standard deviation, for there is a constant relationship between these two. This is expressed by the formula: $P.E. = 0.6745\sigma$.

Relations between Different Measures of Variation

An understanding of the significance of the various measures of dispersion described above may be facilitated by a general comparison and a summary statement of the relations holding between them.

1. The *range* is a distance along the scale within which all the observations lie.
2. The *quartile deviation* or *semi-interquartile range* is a distance along the scale which, when laid off on each side of the point midway between the two quartiles, includes one half the total number of observations.
3. The *mean deviation* from the mean, in a normal or slightly skew distribution, is equal to about $\frac{4}{5}$ of the standard deviation. A range of $7\frac{1}{2}$ times the mean deviation, centering at the mean, will include approximately 99 per cent of all the cases.
4. When a distance equal to the *standard deviation* is laid off on each side of the mean, in a normal or only slightly skew distribution, about two thirds of all the cases will be included;

(In the normal distribution about 68.26 per cent of the observations will be included.) When a distance equal to twice the standard deviation is laid off on each side of the mean approximately 95 per cent of the cases will be included (about 95.46 per cent in a normal distribution). When a distance equal to three times the standard deviation is laid off on each side of the mean about 99 per cent of all the observations will be included (about 99.73 per cent in a normal distribution). This general rule that a range of six times the standard deviation, centering at the mean, will include about 99 per cent of all the measures furnishes a useful check upon calculations.

A study of Fig. 45 may help to make clear the significance of the standard deviation in a normal distribution.

5. The *probable error*, in a normal distribution, is equal to 0.6745σ. A range of twice the probable error, centering at the mean, will include 50 per cent of all the observations. A range of eight times the probable error, centering at the mean, will include approximately 99 per cent of all the observations.

Characteristic Features of the Chief Measures of Variation

The range

1. The range is easily calculated and its significance is readily understood. As a rough measure of the degree of variation the range is useful.
2. The value of the range is determined by the values of the two extreme cases. It is thus a highly unstable measure, the value of which may be greatly changed by the addition or withdrawal of a single figure.
3. This measure gives no indication of the character of the distribution within the two extreme observations.

The quartile deviation

1. The quartile deviation is a measure of dispersion that is easily computed and readily understood. It is superior to the range as a rough measure of variation.
2. The quartile deviation is not a measure of the variation from any specific average.
3. This measure is not affected by the distribution of the items between the first and third quartiles, or by the distribution outside the quartiles. The values of the quartile deviation might be the same for two quite dissimilar distributions, pro-

vided the quartiles happened to coincide. Because it is not affected by the deviations of individual items it cannot be accepted as an accurate measure of variation.

4. The quartile deviation is not suited to algebraic treatment.

The mean deviation

1. The mean deviation is affected by the value of every observation. As the *average difference* between the individual items and the median (or mean) of the distribution it has a precise significance.
2. The mean deviation is less affected by extreme deviations than the standard deviation.
3. Mathematically, the mean deviation is not as logical or as convenient a measure of dispersion as the standard deviation.

The standard deviation

1. The standard deviation is affected by the value of every observation.
2. The process of squaring the deviations before adding avoids the algebraic fallacy of disregarding signs.
3. The standard deviation has a definite mathematical meaning and is perfectly adapted to algebraic treatment.
4. The standard deviation is, in general, less affected by fluctuations of sampling than the other measures of dispersion.
5. The normal curve of error has been analyzed in terms of the standard deviation. The information thus obtained has increased greatly the utility of the standard deviation.

The probable error

1. The probable error has a definite meaning in the case of a distribution following the normal law. It has not this precise meaning for other distributions, and should not be employed in describing them.
2. For distributions to which it is adapted, the probable error is an extremely useful measure. Its most important use is as an index of the magnitude of errors of sampling.
3. The definite relationship between the probable error and the standard deviation, for a normal distribution, permits the value of the probable error to be readily determined.

All the measures of variation described above may be utilized for particular purposes. The standard deviation, however, is the best general measure and should be employed in all cases where a high degree of accuracy is re-

quired. The probable error is, in effect, merely a fractional part of the standard deviation, with a definite but restricted field of usefulness.

The Measurement of Relative Variation

We have been dealing in the preceding section with absolute variability. The various measures of dispersion secured by the methods outlined describe the variability of the data in terms of absolute units of measurement. The standard deviation of London-Paris exchange rates is in francs, the standard deviation of pig iron production in tons, etc. If the object in a given case is the description of a single frequency distribution it is desirable that the original unit be employed throughout, but if measures of variation of two different distributions are to be compared, difficulties are encountered. This is clear if the units are unlike, but even if the units are identical the same difficulty arises. Thus measures of variation in the weights of dogs and in the weights of horses might both have been computed in pounds. Because the standard deviation of horse weights is greater than the standard deviation of dog weights, it does not follow that the degree of variability is greater in the former case. A measure of absolute variation is significant only in relation to the average from which the deviations are measured. Its use, apart from this average, is meaningless. For comparison, therefore, it must be reduced to a relative form, and the obvious procedure is to express a given measure of variation as a percentage of the average from which the deviations have been measured. The quantity thus becomes an abstract number, a measure of the relative variability of the given observations, and may be compared with similar terms computed from other distributions.

THE COEFFICIENT OF VARIATION

The measure of relative variation most commonly employed is that developed by Pearson, termed the *coefficient*

of variation, and represented by the letter V. It is simply the standard deviation as a percentage of the arithmetic mean. Thus

$$V = \frac{\sigma}{M} \times 100.$$

Applying this formula to the results secured from the analysis of the distribution of steel workers, classified according to hourly earnings in 1933 (Table 34), we have

$$V = \frac{18.685}{50.136} \times 100$$
$$= 37.27\%.$$

This measurement may be compared with a similar coefficient relating to the distribution of workers in open-hearth furnaces, classified according to average hourly earnings in 1935. In that year the mean wage was 71.946 cents and the standard deviation 28.55 cents. From these

$$V = \frac{28.55}{71.946} \times 100$$
$$= 39.68\%.$$

Variations of hourly earnings among steel workers was greater in 1935 than in 1933. The difference was not as great, however, as a comparison of standard deviations would indicate. The average wage advanced appreciably between 1933 and 1935 and the relative variation increased only moderately.

An index of variability similar to this coefficient might be secured by expressing any of the other measures of deviation as a percentage of the average from which the deviations were computed. Pearson's coefficient has been generally adopted, however, and is the only one in wide use.

Measures of Skewness

Methods have been developed in the preceding sections for describing the central tendency of a frequency distri-

bution and for measuring the degree of concentration or lack of concentration about that central tendency. One further measure is needed, and that is one which indicates the degree of skewness or asymmetry of a given distribution. For it is essential to know, in regard to a given distribution, whether the observations are arranged symmetrically about the central value, or are dispersed in an uneven, asymmetrical fashion about that value. Having such a figure it will be possible effectively to summarize the characteristics of a frequency distribution in three simple terms — an average, a measure of dispersion and a measure of skewness. There are two measures of skewness in current use.

If a frequency curve is perfectly symmetrical, mean, median, and mode will coincide. As the distribution departs from symmetry these three values are pulled apart, the difference between the mean and the mode being greatest. This difference may be used, therefore, as a measure of skewness. It is desirable in this case, as in measuring relative variability, to secure an index in the form of an abstract number, which may be compared with similar figures derived from other distributions. To this end, Pearson has proposed dividing the absolute difference between mean and mode by the standard deviation of the given distribution. His formula is

$$sk \text{ (skewness)} = \frac{M - Mo}{\sigma}.$$

In a symmetrical distribution, where mean and mode coincide, the value of this measure will be zero. Under other conditions the value may be positive or negative, depending upon the relative positions of the two averages on the scale.

For moderately skew distributions the degree of skewness may be computed more readily from the formula

$$sk = \frac{3(M - Md)}{\sigma}.$$

This corresponds approximately to the other formula, because of the fact that in a moderately asymmetrical distribution the median lies between the mean and the mode, about one third of the distance from the former towards the latter.

Because it is difficult to locate the mode by simple methods, a measure of skewness more easily computed than Pearson's is desirable in some cases. Bowley has proposed such a method, based upon the relationship between the first and third quartiles and the median. If the distribution is symmetrical these two quartiles will be equidistant from the median; with an asymmetrical distribution this is not so. Therefore, if we let q_2 represent the difference between the upper quartile and the median and q_1 represent the difference between the median and the lower quartile, we may use the formula

$$sk = \frac{q_2 - q_1}{q_2 + q_1}$$

as a means of securing a measure of skewness. This value will vary between 0 and $\pm$ 1. For with perfect symmetry $q_2 = q_1$, and the measure is 0; with asymmetry so pronounced that the median and one of the quartiles coincide, either q_2 or q_1 becomes equal to 0, and the formula gives a value of + 1 or − 1. Bowley suggests that a value of 0.1 indicates a moderate degree of skewness, while a value of 0.3 indicates marked skewness.

The values secured from this measure are not, of course, comparable with the values secured from the application of Pearson's formula for measuring skewness.

KURTOSIS

Reference has been made to a fourth measurable characteristic of frequency curves. This is the degree of flat-toppedness, as compared with the normal curve. A measure of *kurtosis*, the technical term for this characteristic, is given in Chapter XIII.

REFERENCES

Bowley, A. L., *Elements of Statistics*, Chap. 6.

Chaddock, R. E., *Principles and Methods of Statistics*, Chap. 9.

Croxton, F. E. and Cowden, D. J., *Practical Business Statistics*, Chap. 10.

Davies, George R. and Crowder, W. F., *Methods of Statistical Analysis in the Social Sciences*, Chap. 4.

Davies, George R. and Yoder, D., *Business Statistics*, Chap. 2.

Day, Edmund E., *Statistical Analysis*, Chap. 11.

King, W. I., *Elements of Statistical Method*, Chaps. 13, 14.

Richardson, C. H., *An Introduction to Statistical Analysis*, Chaps. 4, 5.

Riegel, Robert, *Elements of Business Statistics*, Chaps. 12, 13.

Riggleman, John R. and Frisbee, Ira N., *Business Statistics*, Chap. 9.

Secrist, Horace, *An Introduction to Statistical Methods*, Chaps. 10, 12.

Tippett, L. H. C., *The Methods of Statistics*, Chap. 1.

Waugh, A. E., *Elements of Statistical Method*, Chap. 5.

Yule, G. U. and Kendall, M. G., *An Introduction to the Theory of Statistics*, Chap. 8.

CHAPTER VI

INDEX NUMBERS OF PRICES

The Nature of Index Numbers

The term "index number" has been applied to a number of somewhat similar devices employed in the analysis of statistical series. Index numbers have been most widely used in the study of price changes, but a brief consideration of certain other uses may make clear the essential characteristics of such measures. In its simplest form this name is applied to a term in a time series expressed as a relative number. Thus an index number of cotton consumption in the United States might take the following form:

Table 35

Domestic Cotton Consumption in the United States, 1926–1936

(Consumption in year ended July 31, 1926 = 100)

Year ended July 31	*Cotton consumption (unit: one thousand running bales)*	*Cotton consumption relative*
1926	6,456	100
1927	7,190	111
1928	6,834	106
1929	7,091	110
1930	6,106	95
1931	5,263	82
1932	4,866	75
1933	6,137	95
1934	5,700	88
1935	5,361	83
1936	6,351	98

Similarly the price of a commodity may be expressed as a relative, the price at a given date or for a given period serving as base.

TABLE 36

Average Price of No. 1 Northern Spring Wheat, Minneapolis
1913, 1929–1936

(Average price in year ended June 30, 1913 = 100)

Crop year	*Weighted average price per bushel*	*Relative price*
1913	$0.874	100
1929	1.276	146
1930	0.984	113
1931	0.739	85
1932	0.605	69
1933	0.770	88
1934	1.026	117
1935	1.165	133
1936	1.247	143

The representation of the terms in a time series as relatives, with reference to a fixed base, makes possible a ready comparison of the values for different dates and enables one to follow the trend of the series much more easily than when the data are presented in their original form. Comparison of the trends of different series is also facilitated.

Though the term index number has been applied to such relatives it is better practice to reserve the term for figures which represent the combination of a number of series. The series to be combined may relate to prices, production, consumption, wages, volume of trade, or to any factor subject to temporal variation. (Index numbers have been used also in measuring such geographical differences as arise from variations in living costs from city to city or from country to country.) Quite complex problems may be involved in the construction of any one of these special forms of index numbers, but the essential aim in all cases is to secure a single, simple series that will define the net resultants of the changes occurring in the constituent elements.

A simple index number may be constructed to represent the course of coal and petroleum production in the United States. In the making of such an index it is necessary to

combine in some way production figures for bituminous and anthracite coal and petroleum. The production figures and the corresponding relatives for the three series, from 1922 to 1936, are given in Table 37.

TABLE 37

Production of Bituminous and Anthracite Coal and Petroleum in the United States, 1922–1936

(Production in 1922 = 100)

Year	*Prod. of bit. coal (million sh. tons)*	*Rel.*	*Prod. of anthr. coal (million sh. tons)*	*Rel.*	*Prod. of petrol. (million bbls.)*	*Rel.*
1922	422.3	100	54.7	100	557.5	100
1923	564.6	134	93.3	171	732.4	131
1924	483.7	115	87.9	161	713.9	128
1925	520.1	123	61.8	113	763.7	137
1926	573.4	136	84.4	154	770.9	138
1927	517.8	123	80.1	146	901.1	162
1928	500.7	119	75.3	138	901.5	162
1929	535.0	127	73.8	135	1,007.3	181
1930	467.5	111	69.4	127	898.0	161
1931	382.1	90	59.6	109	851.1	153
1932	309.7	73	49.9	91	785.2	141
1933	333.6	79	49.5	90	905.7	162
1934	359.4	85	57.2	105	908.1	163
1935	372.4	88	52.2	95	996.6	179
1936	434.1	103	54.8	100	1,098.5	197

A rough index of fuel production, based upon these three series, is desired. It is impossible, obviously, to add the original figures, as the units are not the same. This difficulty may be avoided by using the relative figures. A simple average of the three relatives for a given year may serve as the required index. Index numbers thus secured are given in Table 38 on page 164.

In securing this index, by adding the three relative figures for a given year and dividing by three, equal weight has been given to each of the three series. Such an index of

TABLE 38

Index Numbers of Coal and Petroleum Production in the United States, 1922–1936

(Production in 1922 = 100)

Year	*Index*	*Year*	*Index*
1922	100	1930	133
1923	145	1931	117
1924	135	1932	102
1925	124	1933	110
1926	143	1934	118
1927	144	1935	121
1928	140	1936	133
1929	148		

equally weighted relatives has been termed an unweighted index, but the term is misleading. Weights are used, the weights in this case being equal. It is clear that this index based upon equal weights does not reflect faithfully the three series combined in the present instance. For the three series are not of equal importance, as the system of equal weights assumes. The following figures showing the wholesale values in exchange in 1926 of bituminous coal, anthracite coal, and crude petroleum indicate the relative importance of the three series: [1]

Mineral	*Wholesale value in exchange in* 1926
Bituminous coal	$2,157,740,000
Anthracite coal	888,141,000
Petroleum	1,355,989,000

Roughly, these stand to one another in the relation of 5, 2, and 3, and these weights may be assigned to the series under consideration. An index for each year may be computed, using these weights. The example in Table 39, showing the calculations for the years 1922 and 1923, will illustrate the method.

[1] The figures have been compiled by the U. S. Bureau of Labor Statistics.

TABLE 39

Computation of Weighted Index Numbers of Coal and Petroleum Production

Mineral	*Relative production* 1922	*Wt.*	*Wt. × Rel.*	*Relative production* 1923	*Wt.*	*Wt. × Rel.*
Bituminous coal	100	5	500	134	5	670
Anthracite coal	100	2	200	171	2	342
Petroleum	100	3	300	131	3	393
		10	1,000		10	1,405

Index of fuel production, 1922 = 1,000 ÷ 10 = 100
Index of fuel production, 1923 = 1,405 ÷ 10 = 141

The value of the index thus secured for each of the fifteen years covered is shown in Table 40.

TABLE 40

Weighted Index Numbers of Coal and Petroleum Production in the United States, 1922–1936

Year	*Index*	*Year*	*Index*
1922	100	1930	129
1923	141	1931	113
1924	128	1932	97
1925	125	1933	106
1926	140	1934	112
1927	139	1935	117
1928	136	1936	131
1929	145		

Differences between the two series of index numbers are to be expected. The second series, which is the more logically weighted, is, of course, the more accurate of the two, and gives a more faithful representation of the combined effect of the forces affecting the output of coal and petroleum.

Another type of index number is one in which the items in the constituent series are totaled, the aggregate figure, instead of an average, serving as the representative of the entire group. Such a form of index number may be con-

structed only when the different series are all expressed in the same unit. This form is frequently employed as an indication of changes in the level of prices, the aggregate cost of a bill of goods at one period being compared with the aggregate cost of the same goods at other dates. The figures in Table 41 illustrate this type of index.

TABLE 41

Bradstreet's Index of Wholesale Prices in the United States, 1926–1937 [1]

Year	*Index*	*Year*	*Index*
1926	13.02	1932	7.10
1927	12.78	1933	7.86
1928	13.28	1934	9.22
1929	12.67	1935	9.92
1930	10.75	1936	10.10
1931	8.76	1937	11.06

Each of the yearly aggregates quoted above is the sum of the average prices during the year of 96 commodities at wholesale. Before being added all the prices are reduced to the "per pound" basis, so that a certain degree of comparability is secured. Such an index may be readily changed to the relative form, any year being taken as a base and the totals for the other years expressed as percentages of the figure for the base year.

The examples which have been given will indicate some of the many forms which index numbers may take. The term may refer to a simple relative number; it may be applied to an average of relative terms, or to an aggregate of relative or absolute figures. In all the examples given the index has been designed to serve as a measure of change over a period, as an indicator of changes in the values of time series. The term may have a much broader meaning than this. An index of the ability of salesmen might be constructed by giving numerical values to the factors deter-

[1] Construction of this index was discontinued at the end of 1937.

mining their usefulness and securing an average of these values. An index of the efficiency of different departments in a business enterprise might be constructed. In any case, the construction of an index involves the reduction to comparable terms of a number of different factors and the replacement of these several terms by a single figure which may serve as their representative. Comparison is thus facilitated, whether it be comparison over time or space, or comparison with other indices secured by averaging terms relating to a similar unit. In all its forms (except the first limited and exceptional meaning in which it applies to a simple relative) an index number is thus a type of statistical average, and such numbers, in their construction and use, are subject to all the rules and limitations set forth in the development of the subject of averages.

In the present work we are interested only in the application of the index number device to time series. So varied, however, are the rules and practices relating to its application to different types of time series that certain of these types must be treated separately. Our first concern is with index numbers of wholesale prices.

Price Changes

When price movements are surveyed in detail it is difficult to perceive order, or any definite trend. We find a multiplicity of conflicting movements. The price quotations in Table 42 (on page 168), taken at random, are roughly typical of what would be found were the entire field of prices canvassed in order to compare price movements from month to month.

Of the sixteen commodities listed, five showed no price change at all between October and November, 1937, two showed price increases, and in nine cases prices declined. Some of the price movements were inconsiderable, while some marked very material changes. Such, as seen here in miniature, is what happens in the price system as a whole.

TABLE 42

Commodity Prices at Wholesale [1]

Commodity	*Unit*	*Price (wholesale) October, 1937*	*Price (wholesale) November, 1937*
Brick, common building, average of yard prices	1,000	$12.113	$12.113
Pig iron, basic, Valley furnace	Gross ton	23.500	23.500
Cement, Portland, average of plant prices	Bbl.	1.667	1.667
Linseed oil, raw, N. Y.	Pound	.110	.106
Steel billets, rerolling, Pitts.	Gross ton	37.000	37.000
Steel, scrap, Chicago	Gross ton	14.688	12.500
Copper, electrol., refinery	Pound	.119	.108
Lead, pig, N. Y.	Pound	.058	.051
Zinc, pig, N. Y.	Pound	.065	.060
Coal, anthr., chestnut, average of 15 price series, on tracks, destination	Net ton	9.472	9.610
Coal, bit., mine run, average of 27 price series, on tracks, destination	Net ton	4.305	4.303
Crude petroleum, Penn., at wells	Bbl.	2.413	2.350
Gasoline, motor, California, refinery	Gal.	.083	.085
Cotton, middling, N. O.	Pound	.083	.080
Wheat, no. 2 red winter, Chi.	Bu.	1.033	.951
Sugar, granulated, N. Y.	Pound	.048	.048

All prices do not, with absolute uniformity, move up or down or remain constant. Each of the thousands of commodities traded in on the markets of any country, or of the world, moves in its own individual way, subject to a variety of influences. Yet it does not act in isolation. In its price movements it affects other commodities, and is affected by them. And, in addition to the forces peculiar to each commodity, there are broad forces which act throughout the price system, affecting all commodities. It is the busi-

[1] As compiled by the U. S. Bureau of Labor Statistics.

ness of the economic statistician to bring order out of the chaos of price movements taking place at any given time and, out of the multiplicity of minor movements, to pick the broad trends which affect the whole economic system.

The forces bringing about the price movements that are to be studied are numerous and complicated, but some general conclusions may be drawn with regard to them. There are, in the first place, all those changes in production and consumption conditions peculiar to individual commodities and affecting directly the prices of those commodities. The opening of new fields, improvements in production technique in individual cases, changes in fashion and the transfer of demand from some commodities to others, changes in demand and supply with the seasons — all these are causing constant price readjustments. These are the changes which in ordinary times are most obvious, which are brought home directly to the individual merchant or consumer. Such changes affect the whole price system, as has been pointed out, but not in general by causing upward or downward movements in the system as a whole.

These general movements are due to forces that are broader in their scope. The general improvement in production technique and the increase in the productivity of human labor which has resulted have, by increasing the supply of commodities available for consumption, affected prices. Changes in monetary systems and, in particular, changes in the gold supply have exerted a direct and immediate influence upon prices, by affecting the supply of money in circulation. Similar in character have been changes in banking and credit systems and changes in commercial practice that have affected the use of credit instruments and the rapidity of circulation of money and credits. All these forces influence prices, though their incidence is not so specific as are those of the factors affecting individual commodities directly.

Purpose of General Index Numbers of Wholesale Prices

These separate forces cannot be isolated and evaluated. Their joint action causes a perplexing variety of price changes. In studying these changes the problem might be approached from several different points of view. It might be desired to study the readjustments that take place within the price system, to determine the nature and degree of the shifts within the system that come with changing conditions. Such a study would yield valuable information as to the behavior of prices and the character of their interrelations. Our immediate problem, however, is the determination of the net resultant of all these forces. Do all price movements cancel each other so that while some prices move up and some down there is no net change? Or is there at a given time a preponderance of movements in one direction, causing the level of general prices to move upward or downward? If there is such a trend, what is it, and how may it be measured? Are the statistical methods that have been explained in the earlier sections applicable to the solution of this problem?

The first step in this study involves the answering of the last question asked. It has been brought out that methods of summarizing quantitative data have been developed, but that these methods are applicable only when certain conditions are fulfilled. An average, it was noted, has no significance unless it represents a distinct central tendency in a mass of homogeneous data. Moreover, the type of average to be employed depends upon the character of the distribution it is to represent. Until the distribution of the original data is studied no average or other statistical measure can be intelligently employed. We must first, then, determine what the raw materials of the problem are, and study the frequency distributions secured when these raw materials are organized.

For the present a quite general purpose will be assumed, the determination of the change in the level of general wholesale prices between two specific dates. This is equivalent, of course, to measuring the change in the purchasing power of money in wholesale markets. The raw materials of the problem consist of a number of price quotations on individual commodities, quotations being secured for the two dates to be compared. Each pair of quotations measures the change in the price of a single commodity, a change caused by the interplay of many forces. When a great many such price quotations are brought together we have a mass of data representing the interaction of a multitude of forces, some individual and specific in their incidence, some general, affecting the prices of large groups of commodities or of all commodities. What we seek to determine is the net resultant of all these factors. We seek a measure of the composite effect of the numerous forces that are causing individual prices to rise or fall. This measure will constitute an index number of wholesale prices.

The unit with which we must deal is a single price variation. Whether the statistical methods with which we are familiar may be employed in the organization and analysis of a number of such units depends upon the behavior of such units in mass. The following examples illustrate the frequency distributions secured when these data are classified.

Frequency Distributions of Price Ratios

Each price variation is, of course, a ratio, the ratio of the price of a commodity at a given date to the price of the commodity at another date. The ratios may be reduced to a comparable basis by putting them all in the form of relatives, of the type illustrated in the earlier examples of index numbers. Thus, using one of the pairs of price quotations given above, the ratio of the price of steel scrap in November, 1937, to the price in October, 1937, is

$12.500: $14.688, which, in the form of a relative, becomes 85.1:100. In constructing the following frequency table, the prices at wholesale in 1927 of 670 commodities were expressed as relatives, with the 1926 price as a base in each case. The distribution of these 670 relative numbers is shown in Table 43.

TABLE 43

Distribution of the Relative Prices of 670 Commodities in 1927 [1]

(Average prices in 1926 = 100)

Relative prices	*Mid-point* m	*No. of cases* f	*Percentage of total number of cases*
52.5- 57.4	55	1	.1
57.5- 62.4	60	2	.3
62.5- 67.4	65	6	.9
67.5- 72.4	70	7	1.0
72.5- 77.4	75	8	1.2
77.5- 82.4	80	25	3.7
82.5- 87.4	85	50	7.5
87.5- 92.4	90	76	11.3
92.5- 97.4	95	136	20.3
97.5-102.4	100	196	29.3
102.5-107 4	105	83	12 4
107.5-112 4	110	26	3.9
112.5-117.4	115	16	2.4
117.5-122 4	120	14	2.1
122.5-127.4	125	12	1.8
127.5-132.4	130	2	.3
132.5-137.4	135	3	.5
137.5-142.4	140	5	.8
142.5-147.4	145	1	.1
147 5-152.4	150		
152.5-157.4	155	1	.1
		670	100.0

The frequency polygon representing this distribution appears in Fig. 49. For purposes of comparison with similar distributions the figure shows the percentage distribution.

[1] The 670 commodities included were those employed by the U. S. Bureau of Labor Statistics in the construction of its index of wholesale prices. The original figures, and the relatives, appear in *Bulletin 473*, of that Bureau, on "Wholesale Prices, 1913–1927."

The correspondence of this frequency distribution to the standard types portrayed in earlier sections is obvious. There is the same marked concentration about a central tendency, in this case a tendency of prices to remain stable, for 29 per cent of all the cases showed a change not exceeding 2.5 per cent from their prices in the base year. There is also, in this case, a fairly symmetrical distribution about this central tendency, though the range above the mode is

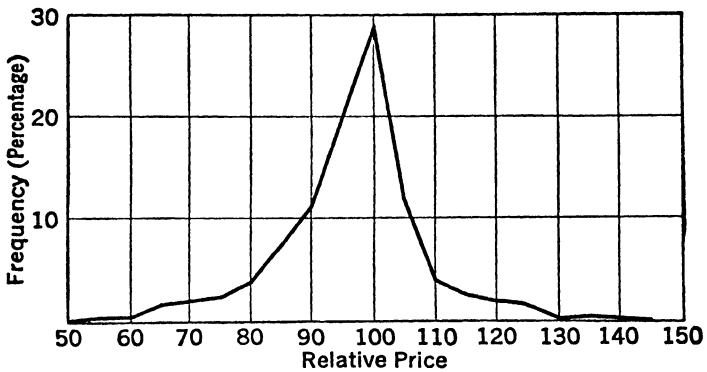


FIG. 49. — Frequency Polygon: Distribution of Relative Prices of 670 Commodities in 1927 (Average prices in 1926 = 100)

slightly greater than the range below. Without at present considering the question as to which average might best be used to represent the central tendency in this distribution, it is apparent that the use of some average is quite legitimate.

The example just given has been based upon price variations from one year to the next, over a period during which the level of general prices declined slightly (4.6 per cent). W. C. Mitchell gives a much more comprehensive illustration, based upon the distribution of 5,578 price variations from one year to the next over the period 1890–1913, which shows the same general grouping. The excess of the range above the mode over the range below is somewhat more pronounced, in connection with which fact it should be noted that prices were rising during most of the 23 years

covered. The distribution secured by Mitchell is shown in Fig. 42.

The inertia of prices is most conspicuous when year-to-year price changes are studied. It is therefore advisable to consider the character of price variations over a longer period, that we may learn whether the same type of distribution is secured. Two examples are given, one of price changes over a seven-year period, marked by a considerable

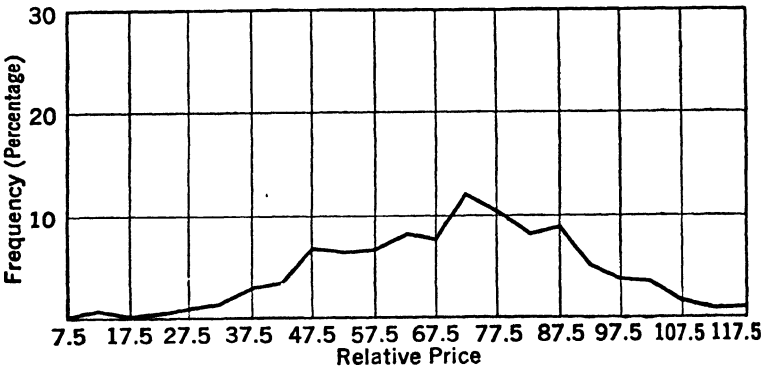


FIG. 50. — Frequency Polygon: Distribution of Relative Prices of 774 Commodities in 1933 (Average prices in 1926 = 100)

decline in prices, the other of price changes over a five-year period characterized by rapidly rising prices. The table following shows the distribution of 774 price variations, prices in 1933 being expressed as relatives on a 1926 base. The general level of wholesale prices, it should be noted, declined some 33 per cent from 1926 to 1933.

The data in Table 44 are plotted in the form of a frequency polygon in Fig. 50, the percentage distribution being shown. It will be noted that the distribution is curtailed, the five upper classes being omitted.

The distributions depicted in Figs. 49 and 50 differ materially. The range of the variations is greater in the second case, a condition naturally to be expected because of the longer period covered. Secondly, a very much smaller percentage of cases is concentrated in the modal group, though

TABLE 44

Distribution of Relative Prices of 774 Commodities in 1933

(Average prices in 1926=100)

Relative prices	*Mid-point* m	*No. of cases* f	*Percentage of total number of cases*
10– 14.9	12.5	3	.4
15– 19.9	17.5		
20– 24.9	22.5	1	.1
25– 29.9	27.5	7	.9
30– 34.9	32.5	13	1.7
35– 39.9	37.5	24	3.1
40– 44.9	42.5	28	3.6
45– 49.9	47.5	51	6.6
50– 54.9	52.5	49	6.3
55– 59.9	57.5	50	6.5
60– 64.9	62.5	62	8.0
65– 69.9	67.5	58	7.5
70– 74.9	72.5	93	12.0
75– 79.9	77.5	81	10.5
80– 84.9	82.5	62	8.0
85– 89.9	87.5	67	8.7
90– 94.9	92.5	40	5.2
95– 99.9	97.5	27	3.5
100–104.9	102.5	27	3.5
105–109.9	107.5	11	1.4
110–114.9	112.5	6	.8
115–119.9	117.5	8	1.0
120–124.9	122.5	1	.1
125–129.9	127.5	2	.3
155–159.9	157.5	1	.1
180–184.9	182.5	1	.1
190–194.9	192.5	1	.1
		774	100.0

there is still a pronounced central tendency. Both distributions, as plotted on the arithmetic scale, are fairly symmetrical, though a few extreme cases extend the actual upper limit of the second distribution. In Fig. 49 the concentration about the central tendency is much more marked, and the deviations of individual price ratios from the central tendency

are smaller. This distribution resembles one which would be secured from highly accurate physical measurements, or the distribution of shots from a very accurate piece of artillery. The second curve corresponds to one representing less accurate physical measurements, or to the distribution of shots from an old or inaccurate field piece. The modal value occurs less frequently and the deviations from the central tendency are greater. It has been established that the longer the period covered in price comparisons such as those made above, the more pronounced is the tendency shown in the second curve. The value of the maximum ordinate falls and the range of the distribution increases. The curve becomes flatter and more extended as the time interval increases. And, quite obviously, as this process goes on the representative character of any type of average declines. Unless there is concentration about a central tendency an average is merely an abstraction, without concrete significance.

It is possible at this point to state as a tentative conclusion that price variations are capable of statistical measurement, that they may be represented appropriately by an average value, provided the period covered is not too long. No definite statement can be made as to the maximum period over which price variations may be measured. Index numbers having accurate and significant values must be based upon comparisons over relatively short periods, the most accurate being year-to-year comparisons. Index numbers designed merely to show general trends in prices may cover longer periods, though the makers and users of such index numbers should realize their limitations.[1]

As a final example we may note the distribution of the relative prices of 1,437 commodities in 1918, average prices during the period July, 1913 to June, 1914 serving as base.[2]

[1] Cf. W. C. Mitchell, "The Making and Using of Index Numbers," *Bulletin 284* (Wholesale Price Series), U. S. Bureau of Labor Statistics.

[2] Data compiled by the Price Section of the War Industries Board; reproduced in Part I, *Bulletin 284*, U. S. Bureau of Labor Statistics, 70.

This was a period marked by rapidly rising prices. In consulting the graph (Fig. 51) it should be noted that the scales are not the same as those employed in the two figures preceding.

A study of this distribution bears out the conclusion reached from the two examples preceding. There is a central tendency sufficiently pronounced to be well represented by an average. In this case, moreover, the modal group is

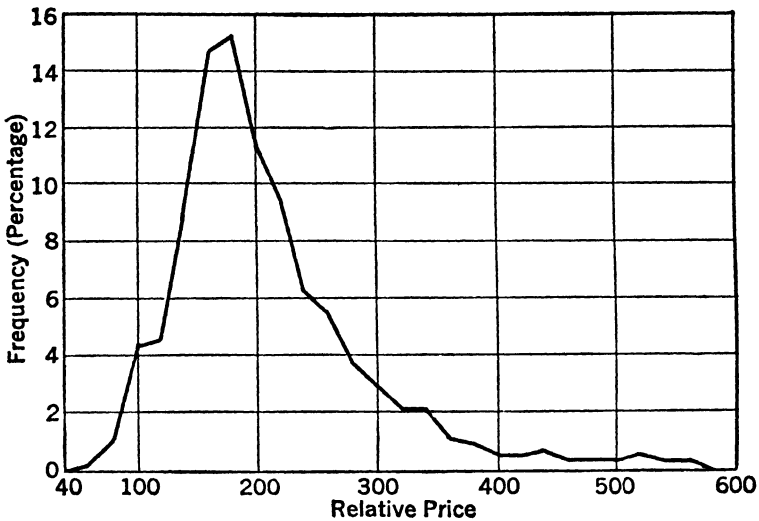


FIG. 51. — Frequency Polygon: Distribution of Relative Prices of 1,437 Commodities in 1918 (Average prices July 1913 to June 1914 = 100)

that with a mid-point of 180, so that the tendency toward concentration cannot be attributed to inertia, but to the presence of external forces affecting the price system as a whole. There is, however, one marked point of difference between this distribution and the two others. The tendency toward skewness, which was in evidence in the first example, is pronounced in this case. The curve, as plotted on the arithmetic scale, is markedly asymmetrical. The greatest concentration is near the lower limit of the scale and a long tail, extending in fact far beyond the limit of the chart,

tapers out to the right. The highest relative price, indeed, is 3,009, representing an increase of 2,909 points. The smallest relative price, in comparison, is 36, representing a decline of 64 points on the scale.

A price increase, expressed as a relative, has no upper limit. An increase of 100, 500, 1,000 per cent or more is conceivable and possible. The greatest price increase noted by the War Industries Board in its study of prices during the war was one of 4,981 per cent, in the case of acetiphenetidin. But 100 per cent is the maximum decline possible, as that would mean that the price of a commodity had fallen to zero. This is the explanation of the skewness noted in the curves shown. When any considerable number of price ratios are tabulated the corresponding frequency curve, plotted on an arithmetic scale, shows this characteristic feature, a feature which is most conspicuous during a period of rising prices.

The argument developed in the preceding pages may be briefly summarized. Before discussing the practice of index number construction it was considered advisable to study the character of the raw materials and the nature of the distributions secured when these materials are brought together, in order to determine whether ordinary statistical methods are appropriate. The raw materials, we have seen, consist of individual price variations, expressed as ratios. When a number of these ratios are assembled a frequency distribution is secured which somewhat resembles the distribution of data following the normal law of error. A central tendency, which may legitimately be represented by an average, is apparent in the distribution of price variations. The central tendency is less marked, however, and the deviations from it are more pronounced, the longer the period covered in the price comparison, so that an average becomes less representative as this period increases. In addition, a tendency toward skewness has been noted, and this was seen to be quite pronounced in a period of rising

TABLE 45

Average Farm Prices, on December 1, of Twelve Leading Crops, 1919–1935 [1]

Crop	*Unit*	1919	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935
Corn	Bu.	1.343	.656	.413	.650	.714	.978	.670	.638	.718	.746	.774	.655	.359	.192	.394	.805	.547
Cotton	Lb.	.356	.139	.162	.238	.310	.226	.182	.109	.196	.180	.164	.095	.057	.057	.094	.124	115
Hay	Ton (sh.)	20.15	17.78	12.09	12.55	14.10	13.80	13 95	14.08	11 30	12.22	12.19	12.62	9.03	6.65	8.10	13.72	7.23
Wheat	Bu.	2.131	1.433	.903	.989	.914	1.309	1.405	1.207	1.118	.981	1.035	.600	.443	.320	.678	.894	894
Oats	Bu.	.702	.456	.298	.390	.408	.476	.374	.392	.443	.403	.426	.315	.230	.134	.304	.525	.257
Potatoes, Wh.	Bu.	1.580	1.128	1.081	.557	.757	.623	1.872	1.413	.951	.527	1.288	.890	.430	.353	.702	.457	.634
Sugar	Lb.*	.102	.053	.037	.057	.073	.053	.041	.051	.046	.039	.038	.033	032	.029	.032	.029	.031
Barley	Bu.	1.215	.716	.421	.525	.535	.747	.586	.571	.675	.547	.544	.389	.353	.201	.407	.778	.377
Tobacco	Lb.	.390	.212	.199	.232	.199	.207	.182	.182	.212	.200†	.186†	.129†	.082†	.105†	.129†	.214†	.185†
Flaxseed	Bu.	4.383	1.770	1.452	2 118	2.107	2.274	2 266	1.941	1.860	2.012	2.843	1.398	1.199	.848	1.518	1.653	1.548
Rye	Bu.	1.331	1 256	.682	.676	.619	1.063	.765	.819	.843	.844	849	384	388	.223	.554	.732	.402
Rice	Bu.	2.666	1.191	.952	.931	1.102	1.386	1.538	1.096	.929	.885	1.002	.782	.608	.391	.779	.792†	.624†

* The price of sugar, given for each year is the wholesale price of the raw product (96° centrifugal) in the month of December at New York. No figure corresponding to the farm price could be secured for this commodity.

† Weighted average price for crop marketing season.

[1] Sources: *Yearbook of Agriculture*, and "Crops and Markets," published by the United States Department of Agriculture. Entries for 1934 and 1935 are averages of farm prices on November 15 and December 15.

Certain of the farm prices here quoted were published by the Department of Agriculture as preliminary figures, subject to later revision.

prices. This skewness is due to the fact that we are dealing with ratios that have a definite lower limit and no upper limit.

Variety of Methods Employed in Index Number Construction

Many methods have been and are being employed in the construction of index numbers of wholesale prices. Usage varies for many reasons. There are differences of opinion as to which is theoretically the best method. There are practical difficulties to be surmounted, difficulties which inevitably cause differences in practice because of the varying resources of the agencies engaged in these tasks. And there are, finally, differences due to the varying purposes for which index numbers are constructed, the varying questions they are designed to answer.

Prevailing differences in practice and differences in the results secured by the employment of various methods in the construction of index numbers can perhaps be illustrated most effectively by the application of a number of methods to the same data. Table 45, on the preceding page, presents the raw material to which these various methods are to be applied — the average farm prices, on December 1, of twelve leading crops, from 1919 to 1935.

Explanation of Symbols

The symbols to be employed in the computation of different types of index numbers have the following meanings:

p_0' : price of a given commodity at time "0" (the base period).
q_0' : quantity of same commodity at time "0".
p_1' : price of same commodity at time "1".
q_1' : quantity of same commodity at time "1".
p_0'' : price of a second commodity at time "0".
q_0'' : quantity of second commodity at time "0".
p_1'' : price of second commodity at time "1".
q_1'' : quantity of second commodity at time "1".
$\frac{p_1'}{p_0'}$: a price relative (relation of price of a given commodity at

time "1" to price of same commodity at time "0").

$\frac{q_1'}{q_0'}$: a quantity relative.

P_0 : price level at time "0".

P_1 : price level at time "1".

Simple Index Numbers of Prices

In his exhaustive analysis of methods of index number construction [1] Irving Fisher distinguishes six fundamental types: the aggregative (or price aggregate), the arithmetic, harmonic, geometric, median, and mode. The latter has never been employed in a practical way, and may be omitted. The characteristics of the five remaining types may be brought out by considering each of them in its simplest form, before examining the more complicated combinations.

AGGREGATES OF ACTUAL PRICES

In the construction of index numbers of the simple aggregative type, commodity prices pertaining to a given date are added; general price changes are measured by comparing the results thus secured for different dates. Using the above symbols

$$\frac{P_1}{P_0} = \frac{\Sigma p_1}{\Sigma p_0}.$$

When such index numbers are constructed from the data of Table 45 the results in Table 46 on page 182 are secured. The actual aggregates are given in column (2); to facilitate comparison the same figures are reduced to relatives, with the 1919 aggregate as base, in column (3).

The results secured by this method of constructing index numbers of prices will be compared shortly with results secured from the same data by other methods. The chief weakness of this type of index number is obvious. This is not an unweighted nor yet an equally weighted index. The influence of each commodity upon the result is dependent upon the price of the unit in which it happens to be

[1] *The Making of Index Numbers*, Houghton Mifflin Co., 1922.

TABLE 46

Index Numbers of Farm Crop Prices

(Aggregates of actual prices)

(1) *Year*	(2) *Index (aggregate of actual prices)*	(3) *Index, relative (1919 = 100)*
1919	$36.349	100
1920	26.790	74
1921	18.690	51
1922	19.913	55
1923	21.838	60
1924	23.142	64
1925	23.831	66
1926	22.499	62
1927	19.291	53
1928	19.584	54
1929	21.339	59
1930	18.290	50
1931	13.211	36
1932	9.503	26
1933	13.691	38
1934	20.723	57
1935	12.844	35

traded. In the present index, hay, which is quoted by the ton, is given more weight than all the other 11 commodities combined, with flaxseed second in importance. The index secured by adding the quotations is weighted in an entirely illogical fashion and cannot be accepted as reflecting the course of farm crop prices.

One method which has been employed for avoiding the unequal weighting caused by the difference in units in which different commodities are traded is to reduce all quotations to the same unit. Thus hay, rice, corn, cotton, and the other commodities might all be quoted by the pound, and these quotations added to secure the index. Yet this method, which has been employed in the construction of Bradstreet's index, merely replaces one system of illogical weighting by an equally illogical one. Equal weight, if such is desired, is not given to all commodities

by this method. Thus, in 1919 hay was worth \$.010075 per pound, cotton \$.356 per pound and rice \$.059 per pound, cotton having a weight in an aggregate of per pound prices 6 times that of rice and 35 times that of hay.

ARITHMETIC AVERAGES OF RELATIVE PRICES

Another method employed in the construction of index numbers involves the reduction of each quoted price to a relative, with reference to the price of the same commodity at a certain basic date, these relative figures then being averaged by any of the conventional methods. The example in Table 47 illustrates the first phase of this process, data for two years being utilized. The year 1919 is taken as base.

TABLE 47

Computation of Relative Prices for the Construction of Index Numbers

(1) *Commodity*	(2) *Unit*	(3) *Price, 1919*	(4) *Relative*	(5) *Price, 1920*	(6) *Relative*
Corn	Bu.	\$ 1.343	100	\$ .656	48.8
Cotton	Lb.	.356	100	.139	39.0
Hay	Ton (sh.)	20.150	100	17.780	88.2
Wheat	Bu.	2.131	100	1.433	67.2
Oats	Bu.	.702	100	.456	65.0
Wh. Potatoes	Bu.	1.580	100	1.128	71.4
Sugar	Lb.	.102	100	.053	52.0
Barley	Bu.	1.215	100	.716	58.9
Tobacco	Lb.	.390	100	.212	54.4
Flaxseed	Bu.	4.383	100	1.770	40.4
Rye	Bu.	1.331	100	1.256	94.4
Rice	Bu.	2.666	100	1.191	44.7
			1,200		724.4

From these figures the arithmetic averages of relative prices in these two years may be readily computed. The formula for any single relative is $\frac{p_1'}{p_0'}$. When there are N relatives the formula for the index number at time "1" is

$$\frac{\Sigma\left(\frac{p_1}{p_0}\right)}{N}.$$

In the present case

$$\text{Index (1919)} = \frac{1{,}200}{12} = 100.$$

$$\text{Index (1920)} = \frac{724.4}{12} = 60.4.$$

Index numbers computed in this way for the years 1919 to 1935, inclusive, are shown in column (3) of Table 50.

This type of index number is usually termed an "unweighted" index of relative prices. It is weighted, however, just as are the types illustrated in the two examples preceding. The quantity employed as weight in each case is the amount of each commodity which would sell for $100 in the base year. In the preceding example the following quantities have been employed as weights:

Corn	74.5 bu.
Cotton	280.9 lbs.
Hay	4.96 tons
Wheat	46.9 bu.
Oats	142.5 bu.
Potatoes	63.3 bu.
Sugar	980.4 lbs.
Barley	82.3 bu.
Tobacco	256.4 lbs.
Flaxseed	22.8 bu.
Rye	75.1 bu.
Rice	37.5 bu.

What has been done, in effect, in the computation of the simple average of relative prices has been to determine the aggregate amount for which the above quantities would sell in each of the eleven years included. At 1919 prices each of the above quantities would sell for $100, the aggregate value being $1,200; at 1920 prices the aggregate value of the above quantities was $724.40. These aggregates, divided by 12, give the index numbers shown in column (3), Table 50: 100 for 1919, 60 (60.4) for 1920, etc. Thus the "unweighted average of relative prices" is in fact a weighted aggregate of actual prices. It is equally weighted in the

sense that the value of the quantity of each commodity employed as weight was equal to $100 in the base year, 1919.[1]

MEDIANS OF RELATIVE PRICES

The median rather than the arithmetic mean may be employed in securing the average of the relative prices for each year. When the relatives in column (6) of Table 47 are arranged in order of magnitude the following distribution is secured:

39.0	58.9
40.4	65.0
44.7	67.2
48.8	71.4
52.0	88.2
54.4	94.4

The smallest relative price is 39.0, the greatest 94.4; the median value is 56.65. This median value is the index number for 1920. All the index numbers computed in this way from the medians of relative prices are presented in column (4), Table 50.

GEOMETRIC AVERAGES OF RELATIVE PRICES

The geometric averages of the relative prices for the various years may now be computed and the results compared with those secured in the preceding examples. A single relative being represented by the symbol $\frac{p_1'}{p_0'}$, the formula for the geometric mean of N relatives is

$$M_g = \sqrt[n]{\frac{p_1'}{p_0'} \times \frac{p_1''}{p_0''} \times \frac{p_1'''}{p_0'''} \times \dots}$$

A geometric mean is generally computed by the aid of logarithms; in this case

$$\text{Log } M_g = \frac{\log\left(\frac{p_1'}{p_0'}\right) + \log\left(\frac{p_1''}{p_0''}\right) + \log\left(\frac{p_1'''}{p_0'''}\right) + \dots}{N}.$$

[1] Attention was called to this characteristic of the simple average of relative prices by F. R. Macaulay, *American Economic Review*, Dec., 1915, 928.

The method of computation may be illustrated for the years 1919 and 1920. The relative prices of the various commodities are repeated from Table 47.

TABLE 48

Computation of Geometric Averages of Relative Prices

(1) *Commodity*	(2) *Relative price,* 1919	(3) *Logarithm of fig. in col.* (2)	(4) *Relative price,* 1920	(5) *Logarithm of fig. in col.* (4)
Corn	100	2.0	48.8	1.68842
Cotton	100	2.0	39.0	1.59106
Hay	100	2.0	88.2	1.94547
Wheat	100	2.0	67.2	1.82737
Oats	100	2.0	65.0	1.81291
Wh. Potatoes	100	2.0	71.4	1.85370
Sugar	100	2.0	52.0	1.71600
Barley	100	2.0	58.9	1.77012
Tobacco	100	2.0	54.4	1.73560
Flaxseed	100	2.0	40.4	1.60638
Rye	100	2.0	94.4	1.97497
Rice	100	2.0	44.7	1.65031
		24.0		21.17231

$$\text{Log } M_g \text{ (1919)} = \frac{24}{12} = 2$$

$$M_g = \text{anti-logarithm of } 2 = 100$$

$$\text{Log } M_g \text{ (1920)} = \frac{21.17231}{12} = 1.76436$$

$$M_g = \text{anti-logarithm of } 1.76436 = 58.1.$$

This value, 58.1, is the index number for 1920. The results for all the years are summarized in column (5), Table 50.

HARMONIC AVERAGES OF RELATIVE PRICES

The characteristics of the harmonic average have been discussed in a preceding chapter. The reciprocal of the harmonic mean, it will be recalled, is the arithmetic mean of the reciprocals of the constituent measures. The constituent items, in the present case, are price relatives of the

form $\frac{p_1'}{p_0'}$. The reciprocal of such a relative is $\frac{p_0'}{p_1'}$. The formula for the harmonic mean of N price relatives is, therefore,

$$\frac{1}{H} = \frac{\frac{p_0'}{p_1'} + \frac{p_0''}{p_1''} + \frac{p_0'''}{p_1'''} + \cdots}{N}$$

or

$$H = \frac{N}{\Sigma\left(\frac{p_0}{p_1}\right)}.$$

The method of computation is illustrated in Table 49.

TABLE 49

Computation of Harmonic Averages of Relative Prices

(1) *Commodity*	(2) *Relative price,* 1919	(3) *Reciprocal of fig. in col.* (2)	(4) *Relative price,* 1920	(5) *Reciprocal of fig. in col.* (4)
Corn	100	.01	48.8	.02049180
Cotton	100	.01	39.0	.02564103
Hay	100	.01	88.2	.01133787
Wheat	100	.01	67.2	.01488095
Oats	100	.01	65.0	.01538462
Wh. Potatoes	100	.01	71.4	.01400560
Sugar	100	.01	52.0	.01923077
Barley	100	.01	58.9	.01697793
Tobacco	100	.01	54.4	.01838235
Flaxseed	100	.01	40.4	.02475248
Rye	100	.01	94.4	.01059322
Rice	100	.01	44.7	.02237136
		.12		.21404998

$$H\ (1919) = \frac{12}{.12} = 100$$

$$H\ (1920) = \frac{12}{.21404998} = 56.1.$$

The index numbers computed in this way for all the years included in the study are shown in column (6), Table 50.

In the construction of the five types of index numbers explained above no attempt has been made to use a logical weighting system. All are termed "unweighted" averages, a term which is quite misleading. The first index constructed, based on aggregates of actual prices, is a heavily weighted index number, though the weights are illogical. In the next four the quantities employed as weights are the amounts purchasable for $100 in 1919. The five results are brought together and compared in Table 50. In each case the index is given to the nearest whole number. These index numbers are plotted in Fig. 52.

Comparison of Simple Index Numbers

The four averages of relative prices agree much more closely with each other than with the index numbers based on aggregates. For reasons already suggested the latter is quite untrustworthy as a measure of price changes. Of the other index numbers, the arithmetic, geometric, and harmonic means show a consistent relationship, a fact which follows from the nature of the averages employed. Except in the base year the geometric mean is always less than the arithmetic and the harmonic is always less than the geometric, the amount of difference increasing as the dispersion of prices becomes greater. The median, with only twelve items to be averaged, is somewhat unstable, and its relationship to the other averages is not always a consistent one.

How are we to choose among these varying results? No one of these "unweighted" index numbers is perfect, for weights which have crept in do not measure the relative importance of the various commodities included in the index numbers. But, neglecting for the moment the question of weights, is it possible to test the adequacy of the different methods of measuring changes in the prices as given?

TABLE 50

Index Numbers of Farm Crop Prices, 1919–1935
(1919 = 100)

(1) Year	(2) *Aggregates of actual prices (as relatives)*	(3) *Arithmetic averages of relative prices*	(4) *Medians of relative prices*	(5) *Geometric averages of relative prices*	(6) *Harmonic averages of relative prices*
1919	100	100	100	100	100
1920	74	60	57	58	56
1921	51	44	42	43	42
1922	55	51	50	50	49
1923	60	55	50	54	53
1924	64	60	61	59	58
1925	66	59	53	57	55
1926	62	53	49	52	50
1927	53	53	55	52	52
1928	54	48	48	47	46
1929	59	54	53	53	52
1930	50	38	32	36	35
1931	36	27	27	27	26
1932	26	20	18	19	19
1933	38	35	33	34	34
1934	57	48	48	46	43
1935	35	35	36	35	34

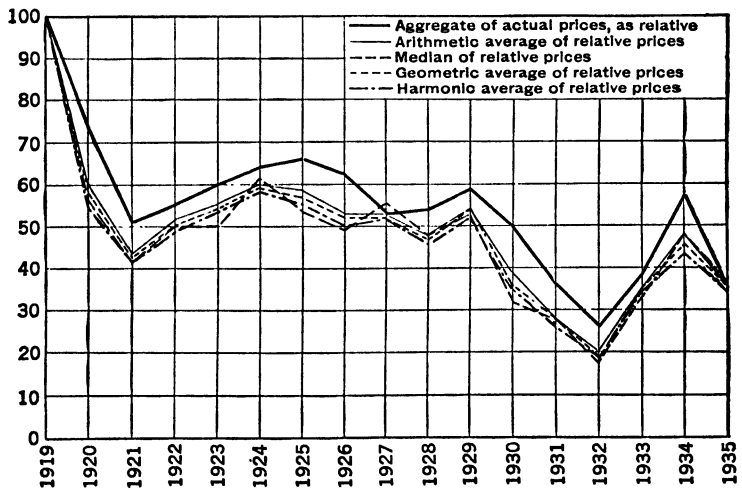


FIG. 52. — Comparison of Five Simple Index Numbers of Farm Crop Prices, 1919–1935 (1919 = 100)

THE TIME REVERSAL TEST

For this purpose Irving Fisher has employed what he terms the "time reversal test." This is merely a test to determine whether a given method will work both ways in time, forward and backward. If from 1935 to 1936 sugar should increase from four to eight cents a pound, the price in 1936 would be 200 per cent of the price in 1935, and the price in 1935 would be 50 per cent of the price in 1936. One figure is the reciprocal of the other; their product ($2.00 \times .50$) is unity. Similarly, if a given method of index number construction shows the general price level in one year to be 200 per cent of the level in the preceding year, it should work correctly when reversed; it should show that the price level in the first year was 50 per cent of the price level in the second year. When the data for any two years are treated by the same method, but with the bases reversed, the two index numbers secured should be reciprocals of each other. Their product should always be unity. If it is not, there is an inherent bias in the method.

This test may be applied to the methods employed above, using prices for 1919 and 1920. With 1919 as base the following results were obtained:

Year	*Aggregates of actual prices (as relatives)*	*Arithmetic averages of relative prices*	*Medians of relative prices*	*Geometric averages of relative prices*	*Harmonic averages of relative prices*
1919	100	100	100	100	100
1920	73.70216	60.36666	56.65	58.1221	56.0617

and with 1920 as base:

Year	*Aggregates of actual prices (as relatives)*	*Arithmetic averages of relative prices*	*Medians of relative prices*	*Geometric averages of relative prices*	*Harmonic averages of relative prices*
1919	135.68122	178.36666	176.85	172.04	165.6467
1920	100	100	100	100	100

When the index numbers for 1920 in the first table are multiplied by the corresponding index numbers for 1919 in the second table, we have the following values. (In securing these products the index numbers are put in the ratio, not in the percentage form.)

Aggregates of actual prices	*Arithmetic averages of relative prices*	*Medians of relative prices*	*Geometric averages of relative prices*	*Harmonic averages of relative prices*
1.00	1.0767	1.00	1.00	.9286

This time reversal test is met by three of the methods employed. It is not met by either the arithmetic or harmonic averages. The former has a distinct upward bias, amounting to more than seven per cent when the errors for 1919 and 1920 are compounded, while the harmonic mean shows almost as large an error in the opposite direction. Unless the inherent bias which is found in both these averages is rectified in some way, methods based upon these averages should not be used in the construction of index numbers.

The Weighting of Index Numbers

Five simple index numbers of prices have been described in the preceding section. With the introduction of weighting the number of possible combinations is greatly increased, but only a few of these types need concern us here.

In the construction of an accurate measure of price changes logical weights must be employed, weights which truly reflect the relative importance of the commodities included. If the weighting problem is ignored haphazard and illogical weights will inevitably be present, whether recognized or not.

The data used in the preceding examples may be utilized to illustrate methods of weighting and to show the effects of varying weights upon the values of index numbers. The

TABLE 51

Annual Physical Production, Twelve Crops, 1919–1935 [1]

Year	*Corn (millions of bu.)*	*Cotton (millions of bales)* [2]	*Hay, tame (millions of short tons)*	*Wheat (millions of bu.)*	*Oats (millions of bu.)*	*White potatoes (millions of bu.)*	*Sugar (millions of lbs.)* [3]	*Barley (millions of bu.)*	*Tobacco (millions of lbs.)*	*Flaxseed (millions of bu.)*	*Rye (millions of bu.)*	*Rice (millions of bu.)*
1919	2,679	11.41	76.59	952.1	1,107	297.3	4,371	131.1	1,444	6.77	78.7	42.69
1920	3,071	13.43	76.16	843.3	1,444	368.9	4,817	171.0	1,509	10.90	61.9	51.65
1921	2,928	7.95	71.04	819.0	1,045	325.3	5,242	132.7	1,005	8.11	61.0	39.27
1922	2,707	9.76	80.79	846.6	1,148	419.3	5,590	152.9	1,254	10.52	101.0	41.66
1923	2,875	10.14	75.29	759.5	1,227	366.4	5,900	159.0	1,518	16.56	56.0	33.24
1924	2,298	13.63	80.12	840.1	1,424	384.8	5,646	167.3	1,245	31.24	59.1	32.59
1925	2,853	16.11	67.16	669.1	1,410	297.6	6,541	192.8	1,376	22.34	42.8	32.74
1926	2,575	17.98	67.48	833.5	1,142	322.4	6,648	164.5	1,289	18.54	35.4	41.42
1927	2,678	12.96	83.65	874.7	1,093	368.8	6,518	240.1	1,211	25.18	52.1	44.42
1928	2,715	14.48	72.59	913.0	1,319	425.6	6,568	329.6	1,373	19.14	38.6	43.43
1929	2,536	14.83	76.11	822.2	1,118	327.7	7,192	280.2	1,537	15.91	35.5	40.60
1930	2,065	13.93	63.57	889.7	1,277	332.7	6,365	303.8	1,647	21.29	46.3	44.92
1931	2,589	17.10	65.34	932.2	1,127	373.0	6,392	198.5	1,607	11.80	32.3	44.87
1932	2,907	13.00	70.20	744.1	1,247	358.0	6,446	302.0	1,023	11.67	40.6	40.41
1933	2,330	13.18	65.85	527.4	722	317.1	6,406	156.1	1,396	6.79	21.2	35.62
1934	1,377	9.64	52.27	496.9	526	385.4	6,278	118.3	1,046	5.21	16.0	38.30
1935	2,203	10.73	75.62	603.2	1,195	356.4		292.2	1,284	14.93	57.9	38.45

[1] Source: *Yearbook of Agriculture*, published by U. S. Department of Agriculture. Certain of the production figures here given were published by the Department of Agriculture as preliminary estimates, subject to later revision.

[2] Bales of 500 pounds gross weight.

[3] The figures for sugar production represent total production of beet and cane sugar in contiguous United States for the crop year indicated, plus imports from non-contiguous United States for the fiscal year beginning July 1 of the crop year, minus domestic exports. Computations for the year 1935 were based on estimated sugar production.

weights employed in constructing index numbers of farm crop prices may be either the quantities or values of the crops produced, depending upon the type of index selected. The quantities produced during the period 1919–1935 are given in Table 51.

WEIGHTED AGGREGATES OF ACTUAL PRICES

The thoroughly illogical results obtained when actual prices, as quoted, are totaled to secure an index number have been pointed out. The same objection cannot be made when the prices are appropriately weighted before the aggregate is taken. If for weights we employ the quantities produced in the base year (at time "0") the formula for the weighted aggregate is

$$\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}.$$

This is, in effect, the method employed by the United States Bureau of Labor Statistics, though the quantities are taken from a year other than the base year. The formula for this type of weighted aggregative index is known as Laspeyres' formula. The method is illustrated in Table 52.

The desired index numbers, in the form of relatives, may be computed from the aggregates secured by totaling columns (5) and (8) of Table 52. Either year may be taken as the base, and the price aggregate in the other year expressed as a relative on this base. With the 1919 aggregate as base the index for 1920 is 58.2. Index numbers similarly computed for the other years are given in column (2), Table 55.

Another type of weighted aggregate may be constructed, with weights taken not from the base period but from the later period in the given comparison. That is, we may employ q_1 (quantity at time "1") as weight in comparing prices at time "1" with prices at time "0," and employ q_2 (quantity at time "2") as weight in comparing prices at

TABLE 52

Computation of Weighted Aggregates of Actual Prices

(1) *Commodity*	(2) *Unit*	(3) *Price, 1919* p_0	(4) *Weight (quantity produced 1919, in millions)* q_0	(5) *Price × weight* p_0q_0	(6) *Price, 1920* p_1	(7) *Weight (quantity produced 1919, in millions)* q_0	(8) *Price × weight* p_1q_0
Corn	Bu.	$ 1.343	2,679	$ 3,597,897,000	$ .656	2,679	$1,757,424,000
Cotton	Lb.	.356	5,705	2,030,980,000	.139	5,705	792,995,000
Hay	Ton (sh.)	20.150	76.59	1,543,288,500	17.780	76.59	1,361,770,200
Wheat	Bu.	2.131	952.1	2,028,925,100	1.433	952.1	1,364,359,300
Oats	Bu.	.702	1,107	777,114,000	.456	1,107	504,792,000
Potatoes, Wh.	Bu.	1.580	297.3	469,734,000	1.128	297.3	335,354,400
Sugar	Lb.	.102	4,371	445,842,000	.053	4,371	231,663,000
Barley	Bu.	1.215	131.1	159,286,500	.716	131.1	93,867,600
Tobacco	Lb.	.390	1,444	563,160,000	.212	1,444	306,128,000
Flaxseed	Bu.	4.383	6.77	29,672,910	1.770	6.77	11,982,900
Rye	Bu.	1.331	78.7	104,749,700	1.256	78.7	98,847,200
Rice	Bu.	2.666	42.69	113,811,540	1.191	42.69	50,843,790
				$11,864,461,250			$6,910,027,390

time "2" with prices at time "0." Algebraically, the formula for the index number at time "1" is

$$\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}.$$

This is known as Paasche's formula. The process of computation is precisely the same as in the preceding example, except that the weights are changed with each successive year. The index numbers secured by this method are given in column (3), Table 55.

The weights in these two cases have been *quantities*, for prices, multiplied by quantities, give aggregates in dollar values. But in weighting individual price relatives, quantities will not serve. The abstract relatives must be weighted by *values*, if the resulting products are to be comparable. For values are in terms of a common dollar unit, while quantities may be expressed in a variety of units. The values which are to be employed as weights may be derived in various ways.

Fisher [1] outlines the four following methods, of which the second and third are hybrid types:

I. Each weight = base year price × base year quantity (p_0q_0).
II. Each weight = base year price × given year quantity (p_0q_1).
III. Each weight = given year price × base year quantity (p_1q_0).
IV. Each weight = given year price × given year quantity (p_1q_1).

Just as certain averages possess inherent bias, so a distinctive weight bias arises from each type of value weighting. (This inherent bias is absent from the quantity weighting.) A downward bias arises from weighting systems I and II (in which base year prices are used), while an upward bias arises from weighting systems III and IV (using prices in the given year). This is in part capable of mathematical demonstration [2] and has in part been established by numerous trials.

[1] Irving Fisher, *The Making of Index Numbers*, 54.

[2] An index weighted by type III must exceed an index weighted by type I.
(*Footnote 2 continued on page 196.*)

In the several examples next following we shall deal only with values of quantities produced in the base year, 1919. These values are given in the third column of Table 53. For weighting purposes they are taken to the nearest million.

WEIGHTED ARITHMETIC AVERAGES OF RELATIVE PRICES

In the computation of an index of this type, each relative is multiplied by the appropriate weight and the sum of the products is divided by the sum of the weights. The process is illustrated in Table 53.

The index for 1920, it will be noted, is identical with that secured from the computations illustrated in Table 52. That index is a weighted aggregate of actual prices, the weights being the quantities produced in the base year. An arithmetic mean of relative prices, weighted by values in the base year, is always equal to a relative constructed from such an aggregate.[1]

(*Footnote 2 continued from page 195.*)
Weighting the price relative of a given commodity by type III, we have

$$\frac{p_1}{p_0} \times p_1 q_0$$

while by type I we have

$$\frac{p_1}{p_0} \times p_0 q_0.$$

If p_1 exceeds p_0 (if the price relative is above 100) the weight by type III ($p_1 q_0$) is greater than the weight by type I ($p_0 q_0$). That is, all relatives above 100 are more heavily weighted by type III than by type I. But if p_1 is less than p_0 the weight by type III ($p_1 q_0$) is less than the weight by type I ($p_0 q_0$). All relatives below 100 are less heavily weighted by type III than by type I. Thus the effect of all price increases is over-emphasized and the effect of all price declines is under-emphasized by type III, giving a net result always greater than type I. The same is true of type IV as compared with type II. As between types I and IV there is no necessary relation, but in general an index weighted by type IV will exceed an index weighted by type I. Base year weighting involves a downward bias while given year weighting involves an upward bias. (For a more detailed discussion of bias in weighting see Fisher, *The Making of Index Numbers*, Chapter V and pages 384–387.)

[1] This may be readily demonstrated algebraically. The value of any commodity in the base year is $p_0 q_0$, while the price relative for a second year is $\frac{p_1}{p_0}$.

(*Footnote 1 continued on page 197.*)

TABLE 53

Computation of Weighted Arithmetic Averages of Relative Prices

Commodity	*Relative price, 1919*	*Weight*	*Relative price × weight*	*Relative price, 1920*	*Weight*	*Relative price × weight*
Corn	100	3,598	$359,800	48.8	$3,598	$175,582 4
Cotton	100	2,031	203,100	39.0	2,031	79,209.0
Hay	100	1,543	154,300	88.2	1,543	136,092.6
Wheat	100	2,029	202,900	67.2	2,029	136,348.8
Oats	100	777	77,700	65.0	777	50,505.0
Potatoes	100	470	47,000	71.4	470	33,558 0
Sugar	100	446	44,600	52.0	446	23,192.0
Barley	100	159	15,900	58.9	159	9,365.1
Tobacco	100	563	56,300	54.4	563	30,627.2
Flaxseed	100	30	3,000	40.4	30	1,212 0
Rye	100	105	10,500	94.4	105	9,912.0
Rice	100	114	11,400	44.7	114	5,095.8
		$11,865	$1,186,500		$11,865	$690,699.9

(The weights employed are the values of the quantities produced in 1919, in millions.)

$$\text{Weighted arithmetic mean (1919)} = \frac{\$1,186,500}{\$11,865} = 100$$

$$\text{Weighted arithmetic mean (1920)} = \frac{\$690,699.9}{\$11,865} = 58.2$$

(*Footnote 1 continued from page 196.*)
The weighted mean of such price relatives is equal to

$$\frac{\frac{p_1'}{p_0'} \times p_0'q_0' + \frac{p_1''}{p_0''} \times p_0''q_0'' + \frac{p_1'''}{p_0'''} \times p_0'''q_0''' + \dots}{p_0'q_0' + p_0''q_0'' + p_0'''q_0''' + \dots}$$

which reduces to

$$\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0},$$

a weighted aggregate of the type mentioned.

In the same way the harmonic mean, weighted by full values in the second year, reduces to

$$\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}.$$

This has already been encountered as an aggregate of actual prices weighted by quantities in the second year.

WEIGHTED GEOMETRIC AVERAGES OF RELATIVE PRICES

The process of computing the weighted geometric mean is identical with that of computing the unweighted geometric mean, except that the logarithm of each relative is multiplied by the given weight and the sum of these weighted logarithms is divided by the sum of the weights, the result being the logarithm of the desired index.[1] The method is illustrated in Table 54.

TABLE 54

Computation of Weighted Geometric Average of Relative Prices, 1920
(1919 = 100)

Commodity	*Relative price*, 1920	*Logarithm of relative price*	*Weight*	*Logarithm of relative price × weight*
Corn	48.8	1.68842	3,598	6074.93516
Cotton	39.0	1.59106	2,031	3231.44286
Hay	88.2	1.94547	1,543	3001.86021
Wheat	67.2	1.82737	2,029	3707.73373
Oats	65.0	1.81291	777	1408.63107
Potatoes, Wh.	71.4	1.85370	470	871.23900
Sugar	52.0	1.71600	446	765.33600
Barley	58.9	1.77012	159	281.44908
Tobacco	54.4	1.73560	563	977.14280
Flaxseed	40.4	1.60638	30	48.19140
Rye	94.4	1.97497	105	207.37185
Rice	44.7	1.65031	114	188.13534
			11,865	20,763.46850

$$\text{Log } M_g = \frac{20{,}763.46850}{11{,}865} = 1.74998,$$

$$M_g = 56.2$$

The index for 1920 on the 1919 base is 56.2. Measurements secured for all the years of the period covered are given in column (5), Table 55, together with the other weighted index numbers already explained.

[1] The formula for the weighted geometric mean is given in Chapter IV.

How are we to judge of the relative merits of these three index numbers? We may, first, apply the time reversal test which was employed in comparing the five simple index numbers. This test is not met by any of the weighted types we have constructed. The geometric is equally at fault with the others. Though the simple geometric meets the test, the introduction of weighting imparts a bias to the result. Judged by that test alone none of the three is satisfactory. We may next try the second fundamental test that Fisher has developed, which is termed the "factor reversal test."

THE FACTOR REVERSAL TEST

The total value of a given commodity in a given year is, of course, the product of the quantity produced and the price per unit; algebraically, it is equal to $p'q'$. The ratio of the total value in one year to the total value in the preceding year is $\frac{p_1'q_1'}{p_0'q_0'}$. If, from one year to the next, both price and quantity should double, the price relative would be 200, the quantity relative 200, and the value relative 400. The total value in the second year would be four times the value in the first year. The value relative would be equal to the product of the price and quantity relatives, a relationship which is obvious in the case of a single commodity.

If, for a number of commodities, we construct an index of the price change from one year to the next and an index of the quantity change from one year to the next, we should expect their product to be equal to the ratio of the total values in the second year to the total values in the first year. If the product is not equal to the value ratio, there is an error in one or both of the index numbers.

As an illustration, we may apply this test to the first aggregative index constructed $\left(\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}\right)$. An index of quantities may be computed from this same formula, merely

interchanging the q's and the p's; the formula becomes

$$\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0}.$$

The same price factor appears in numerator and denominator, as we desire to measure only the effect of the quantity change. Substituting the given values of the twelve farm crops we have

$$\text{Quantity index, 1920 (1919 = 100)} = \frac{\$12,998,610,800}{\$11,864,461,250} = 1.0956.$$

In percentage form the index of quantities produced in 1920 is 109.56, with 1919 as base. The corresponding price index, by the same formula, is 58.24. The product

$$1.0956 \times .5824 = .6381.$$

That is, if prices have decreased 41.76 per cent, while quantities have increased 9.56 per cent, the total value should show a decrease of 36.19 per cent.

For the value ratio we have

$$\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} = \frac{\$7,441,317,450}{\$11,864,461,250} = .6272.$$

There is a discrepancy here of about one per cent. The actual error is not great, but the formula definitely fails to meet the factor reversal test, and cannot be accepted as satisfactory.

When this test is applied to the second aggregative index we secure the following values for 1920, with respect to 1919 as base:

$$\text{Price index} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} = 57.25$$

$$\text{Quantity index} = \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1} = 107.69$$

$$\text{Product} = .5725 \times 1.0769 = .6165$$

(In securing the product the index numbers are put in the ratio, not in the percentage form.)

Here is an error of the same magnitude in the other direction.

The weighted geometric average also fails to meet this fundamental factor reversal test. With respect to both the geometric index and the aggregates we have, apparently, by the introduction of weights spoiled index numbers which in their simple form were unbiased. Yet weights we must have, if the index numbers are to represent the facts accurately. Neither a simple index nor a weighted form of a simple index will meet the two tests laid down as fundamental. Professor Fisher tested 46 such formulas, of which only four (the simple geometric, median, mode, and aggregative) met the time reversal test, and none met the factor reversal test.

THE "IDEAL" INDEX

A way out of this difficulty is offered by the possibility of "rectifying" formulas in a crossing process, by averaging geometrically formulas which err in opposite directions. Professor Fisher has made exhaustive trials of all possible formulas by this process, finding thirteen formulas in all which met both tests. Of these he has selected one as "ideal," from the viewpoint of both accuracy and simplicity of calculation. This ideal index is the geometric mean of the two aggregative types illustrated above. Its formula [1] is

$$\sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}.$$

This index may be computed readily, in the present instance, from the results already obtained. Thus for 1920 we have

$$\text{Ideal index} = \sqrt{.5824 \times .5725}$$
$$= .5774.$$

In the customary percentage form this is 57.74.

This index number meets both the time reversal and the factor reversal test. Applying the former:

[1] The same formula was developed independently by Bowley, Pigou, Walsh, and Young. See *The Making of Index Numbers*, xv, 240–242.

Index of prices, 1920 (1919 = 100) = 57.74
Index of prices, 1919 (1920 = 100) = 173.18
$.5774 \times 1.7318 = 1.00.$

For the factor reversal test, applied to the data for 1920 (with 1919 as base), we have

$$\text{Index of prices} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} = 57.74.$$

$$\text{Index of quantities} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} = 108.62.$$

$$\text{Value ratio} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} = .6272.$$

Product of price and quantity indices $= .5774 \times 1.0862 = .6272.$

The ideal index, the two weighted aggregates that enter into its construction and the geometric mean weighted by

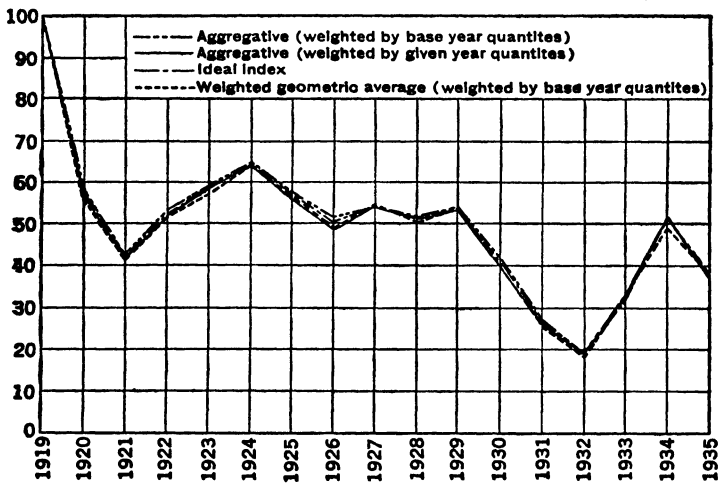


FIG 53. — Comparison of Four Weighted Index Numbers of Farm Crop Prices, 1919–1935 (1919 = 100)

values in the base year are given in Table 55 for the years 1919 to 1935. The index numbers are plotted in Fig. 53.

The wide discrepancies that were found between the various simple index numbers do not appear when the weighted

TABLE 55

Comparison of Weighted Index Numbers of Farm Crop Prices, 1919–1935

(1)	(2)	(3)	(4)	(5)
Year	*Aggregative (weighted by base year quantities)* $\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}$	*Aggregative (weighted by given year quantities)* $\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}$	*Ideal index Geometric mean of indices in cols. (2) and (3)*	*Weighted geometric average (weighted by base year quantities)*
1919	100.0	100.0	100.0	100.0
1920	58.2	57.2	57.7	56.2
1921	42.8	42.0	42.4	41.5
1922	53.6	53.1	53.4	52.9
1923	59.8	59.7	59.8	58.1
1924	65.0	64.3	64.6	64.4
1925	57.9	56.3	57.1	56.5
1926	51.4	49.2	50.3	49.6
1927	54.5	54.3	54.4	54.3
1928	51.8	51.1	51.4	51.2
1929	54.1	53 3	53.7	53.4
1930	41.3	39.6	40.4	39.4
1931	26.6	25.5	26.0	25.3
1932	19.0	18.9	19.0	18.1
1933	32.6	32.2	32.4	32.2
1934	51.6	52.1	51.8	49.5
1935	38.1	37.6	37.9	37.8

indices are compared. There are significant differences, but there is none of the erratic behavior of some of the simpler forms.

Of these four types the ideal index probably serves as the best measure of the average price change between 1919 and each of the given years.[1] It is designed, it should be remembered, to measure the change between two stated times, and not for intermediate comparison. The value of the index for 1933, for instance, is determined by the relation between prices and quantities in 1919 and in 1933.

[1] The year 1919, which is here employed as base, is not a satisfactory standard of reference for economic purposes. It was a disturbed year, marking a transition from war-time to peace-time conditions.

There is double weighting and the weights vary from year to year. If 1933 is to be compared with 1932 a new index is needed, in which the prices and quantities for 1933 and 1932 alone are included. Direct comparison on the basis of the values for the ideal index given in the above table is liable to error, because of the weighting system employed.

It is one of the merits of the geometric mean with constant weights that it permits the index for each year to be compared directly not only with the base year index, but with the index for any other year. The base may be shifted directly from the relatives, and the same result will be secured as if the computation were made from the original data. If this same system be followed with the ideal index no large errors may be expected, but strict accuracy will not be secured.[1]

SOME ALTERNATIVE TYPES

The chief obstacles in the way of general adoption of the ideal index arise from the difficulty of obtaining annual or monthly quantities to use as weights, and from the time involved in its computation. Where accuracy is essential the latter is not a serious difficulty. As a substitute formula which is much more quickly calculated Fisher has proposed

$$\frac{\Sigma(q_0 + q_1)p_1}{\Sigma(q_0 + q_1)p_0}.$$

This formula, which has also been recommended by Edgeworth and Marshall, is considered by Fisher to be "the best practical all-around formula, taking all four points into account — accuracy, speed, minimum legitimate circular discrepancy, simplicity." Results from this formula will generally differ from those secured from the ideal for-

[1] If year to year comparison be a primary aim in a given instance, the ideal index may be constructed on the chain system. Link index numbers are first constructed, each year serving as base for the computation of the index for the succeeding year. These links may then be "chained" with reference to a fixed base. Warren M. Persons has shown that the errors involved in following this method are cumulative, and may be serious if the links are chained for a number of years.

TABLE 56

Computation of Aggregative Index, Weighted by Combined Quantities

(1) *Commodity*	(2) *Unit*	(3) *Price,* 1919	(4) *Quantity* 1919 + *quantity* 1920 (*in millions*)	(5) *Price* 1919 × *sum of quantities* [*col.* (3) × *col.* (4)]	(6) *Price,* 1920	(7) *Price* 1920 × *sum of quantities* [*col.* (6) × *col.* (4)]
Corn	Bu.	$ 1.343	5,750	$ 7,722,250,000	$ .656	$ 3,772,000,000
Cotton	Lb.	.356	12,420	4,421,520,000	.139	1,726,380,000
Hay	Ton (sh.)	20 15	152.75	3,077,912,500	17.78	2,715,895,000
Wheat	Bu.	2 131	1,795.4	3,825,997,400	1.433	2,572,808,200
Oats	Bu.	.702	2,551	1,790,802,000	.456	1,163,256,000
Potatoes, Wh.	Bu.	1.580	666.2	1,052,596,000	1.128	751,473,600
Sugar	Lb.	.102	9,188	937,176,000	.053	486,964,000
Barley	Bu.	1.215	302.1	367,051,500	.716	216,303,600
Tobacco	Lb.	.390	2,953	1,151,670,000	.212	626,036,000
Flaxseed	Bu.	4.383	17.67	77,447,610	1.770	31,275,900
Rye	Bu.	1.331	140.6	187,138,600	1.256	176,593,600
Rice	Bu.	2.666	94.34	251,510,440	1.191	112,358,940
				$24,863,072,050		$14,351,344,840

$$\frac{\Sigma(q_0 + q_1)p_1}{\Sigma(q_0 + q_1)p_0} = \frac{\$14,351,344,840}{\$24,863,072,050}$$

$$= 57.7 \text{ (index for 1920 on 1919 base, in percentage form)}$$

mula by less than one fourth of one per cent. Table 56 on page 205 illustrates the method of computation, data for 1919 and 1920 being employed.

This formula requires the same data as the ideal index, and these are not generally to be had. Usually it is only possible to secure comprehensive quantity figures at each census period, and for the intervening years constant weights must be employed. In such cases the weighted aggregative

$$\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}$$

is probably the most generally useful type. The weighted geometric has many virtues, but is subject to a definite weighting bias. If no weights can be secured, or even approximated, the simple geometric and the simple median are far better than any of the other simple types. The geometric mean is more generally useful than the median.

An index number of prices is always based upon the study of a sample, the result being taken as representative of the entire field of prices from which the particular sample was drawn. Some method is needed, therefore, by which we may judge of the reliability of the different types of index numbers, of their probable stability when computed from a number of successive samples. Some differences might be expected between index numbers based upon different samples. With which type of index number would these differences due to fluctuations of sampling be least? [1]

Truman L. Kelley [2] has attempted to measure the probable errors of the chief types of index numbers and has graded these types on the basis of excellence in this respect. Two index numbers, the weighted geometric mean and the weighted median, are given the highest grade, as being the most reliable, the least affected by fluctuations of sampling.

[1] The subject of sampling, in relation to the reliability of statistical measures, is discussed in greater detail below.

[2] Truman L. Kelley, *Statistical Method*, New York, Macmillan, 1921, 334–346.

Fisher's ideal index is ranked somewhat lower, though above the weighted arithmetic and harmonic averages of price relatives. The simple unweighted arithmetic average of relatives is given the lowest rating in the list.

For reliability, flexibility, and general excellence Kelley selects the weighted geometric mean as the best type of price index number. A ratio of aggregates

$$\frac{\Sigma p_1 w}{\Sigma p_0 w}$$

with selected weights (not necessarily precisely equal to the quantities marketed or consumed) is given a total score, based on the essential requirements of a good index number, as high as that of the weighted geometric mean and higher than that of the ideal index. Weights other than actual quantities are used in order that there may be flexibility in the matter of weighting.

The detailed discussion of procedures in the preceding pages has clearly shown that there are some definitely faulty formulas, obviously unsuited for use in the construction of index numbers serving ordinary purposes. Among the better formulas there are some differences in respect of liability to bias and character of data needed, and some variations in sampling reliability. The maker of index numbers will have these in mind in choosing a formula to employ under given conditions. A more important factor in his choice, however, will be the purpose to be served by the index number, the question it is designed to answer. A weighted aggregate of actual prices answers one question definitively. It gives, without equivocation, the aggregate cost of a fixed bill of goods at one period, in relation to the cost of the same bill of goods at another. A geometric mean of relative prices answers another question. It measures with accuracy the average *ratio* of the prices of given commodities at one period to corresponding prices at another period. Some questions (for example, that answered by an unweighted arithmetic average of relative prices) have little if any economic sig-

nificance. It is because one or two main questions have bulked large in economic discussion that emphasis has been placed upon the finding of a "best" type of index number. Yet the terms "best" and "ideal" are unfortunate, for they imply that some absolute standard exists, with reference to which all formulas may be tested. No such absolute criterion may be applied to the diversity of research problems that call for the construction of index numbers. On the basis of his knowledge of the characteristics of different formulas, the discriminating investigator will choose technical methods adapted to his data and appropriate to his purposes.

Other Problems Involved in the Construction of Price Index Numbers

The preceding section has dealt with the technical problems connected with the averaging of a given set of data in order to secure an index number of price variations. Certain methods have been shown to be quite faulty, while certain others have been found to be appropriate for given purposes. One who would use index numbers with intelligence should understand fully the methods which have been employed in securing given results, in order that he may know precisely what the given figure is designed to measure and what degree of reliability attaches to it.

Such problems as these are not the only ones which confront those who construct index numbers, nor are these considerations the only ones which users of index numbers should bear in mind. Of equal importance with problems of averaging and weighting are the practical questions connected with the selection of representative samples. The only completely accurate measure of the general level of commodity prices would be secured by determining the ratio between all money units, including credit, in circulation (account being taken of velocity of circulation) and all the physical units of goods exchanged for money over a given period. The measurement of general price changes between

two periods would thus involve complete knowledge of these two factors for each of the two periods. Such knowledge, of course, cannot be had, so recourse must be had to the method of sampling. And primary importance attaches to the number of commodities and the character of the commodities upon the prices of which a given index number is based.

NUMBER OF COMMODITIES TO BE INCLUDED

Here again we are confronted with a relation that has already been mentioned, the relation between methods and uses. Decision as to the number of commodities and the kinds of commodities to be included in a given case must rest upon the purpose for which the index is to be constructed. Assuming that the index number is to serve as a measure of general changes in the price level, the question as to the number of commodities to be included may be easily answered — the larger the sample the more representative will be the results. The frequency polygon based upon a large sample will approach more closely to the ideal curve which would represent all price quotations than will that based upon a small sample. Thus, as a measure of general price changes, more confidence may be placed in the Bureau of Labor Statistics index, which is based upon 813 price quotations, than in Bradstreet's, which was based upon 96 quotations, though the latter had particular virtues of its own.[1] Yet index numbers based upon a small number of quotations may not be ruled out as without value. Wesley C. Mitchell, whose researches have materially increased our knowledge of the price system and of the characteristics of index numbers, has compared in detail index numbers based upon varying numbers of quotations. Unexpected similarities are found. Those constructed from a limited number of quotations reflect the broad movements of prices in much the same way as do those based upon the

[1] Bradstreet's index was discontinued at the end of the year 1937.

prices of several hundred commodities. In important details there are differences, however, differences which may involve doubt as to the movement of prices in a given year. In such cases the index numbers based upon many quotations must be accepted as more accurate measures of general price movements, provided that the commodities included be equally representative of the various elements of the price system.

For other purposes, however, index numbers based upon a limited number of quotations may be preferable. This is particularly true when a "sensitive" index is desired, one that will serve as a forecaster of general price movements rather than as a precise measure of changes in the general price level. Of this type is the Harvard sensitive price index based upon quotations on 13 basic commodities (raw materials). The purposes of such an index are served by the selection of a limited number of commodities the prices of which are subject to extreme fluctuations, rather than by the inclusion of a great many commodities. Yet the uses to which an index of this type may be put are limited. The "sluggishness" of the many-commodities index number is a sluggishness which inheres in the price system, and which must be reflected in a faithful index of general prices.

The question of the number of commodities to be included cannot be discussed apart from that of the character of these commodities. The representative character of an index number rests in part upon the number of price series included, but the nature of these series is of even greater importance. For there are highly significant differences in the behavior of the prices of different commodity groups. These groups of prices, their interrelations, their behavior, their relation to the functioning of the economic system and to the swings of prosperity and depression, are matters of immediate and practical importance to economists and business men.

PRICE GROUPS IN THE FIELD OF WHOLESALE PRICES

Since an index number of wholesale prices must rest upon sample quotations, the sample must be representative, must include commodities whose prices are typical of the various elements in the price system. The division into elements for this purpose must be based upon the character of the price changes peculiar to the different groups. Of the groups thus distinguished, the most obvious are those representing different industries. Textile prices and steel prices, leather prices and the prices of chemicals are subject to different influences. Trade depressions and revivals do not affect all industries at the same time or in the same way, so that an index of wholesale prices must include quotations from all important industrial groups. If preponderant influence upon an index is exerted by the prices of certain types of commodities, the index, by that much, loses its representative character. Thus Bradstreet's index, it has been established, gave greater weight to cotton fabrics, hides and leather, and cured meats than was justified by their actual importance in trade, a fact which did not detract from its utility for some purposes but which lessened its value as a representative index of wholesale prices.

The extent of these differences between the price movements of commodities in different industrial groups may be appreciated by comparison of the index numbers of wholesale prices of grains and metals and metal products during the business recession that began in the summer of 1937.

In order that an index may be representative it is not alone sufficient that all industries be given an appropriate number of representatives in the sample. Raw materials and manufactured goods show characteristic differences in their fluctuations, and fitting representation must be given to each of these groups. Prices of the former are, in general, more sensitive to changes in business conditions, their movements preceding those of manufactured goods and showing more

violent fluctuations. There are several reasons for this. Raw materials are traded in for purposes of manufacture and sale. When business improves after a period of depression, increased demand on the part of consumers (or expected increase in demand) leads competing manufacturers to bid against each other for raw materials. It is in the raw material markets that the pressure of increased demand first centers, and this bidding generally causes prices to rise in these markets before the prices of other goods are affected. Similarly, at the first evidence of slackening trade manufacturers' demand for raw materials falls off. Business forces pure and simple play in the raw material markets with more freedom than in the markets for manufactured goods. Hence the tendency of prices in these markets to anticipate, in their movements, prices in other commodity markets.

Additional reasons for the greater stability of prices of manufactured goods are found in the fact that these prices include a greater percentage of stable cost factors, and in the control over supply exercised by most manufacturers. Wages, interests, rents move more slowly and less violently than do commodity prices. The inclusion of these elements in commodity prices tends to render these prices more stable. Therefore, as commodities move forward from the raw stage to their final manufactured condition their prices include more and more of these stabilizing elements, and become less violent in their fluctuations.[1] Control over supply, which manufacturers possess in much higher degree than primary producers, makes possible the enforcement of definite price policies by fabricators. Under these conditions, stable prices and variable output are usually found.

Each of the groups last mentioned contains minor groups of commodities with distinct price characteristics. Within the raw material group there are marked differences between

[1] Cf. Mitchell, "The Making and Using of Index Numbers," *Bulletin 284* U. S. Bureau of Labor Statistics, 44–45, for examples.

agricultural products, animal products, forest products, and mineral products. Agricultural products are affected by weather and crop conditions as well as by business conditions and, though subject to price fluctuations of some magnitude, reflect prevailing business conditions less accurately than do the prices of mineral products.[1] Animal and forest products appear to stand between these two with respect to the faithfulness with which they reflect business conditions in their price movements. Thus, in selecting raw materials for inclusion in a sample of price quotations from which a representative index number is to be constructed, fair weight must be given to these various classes.[2]

Manufactured goods, again, do not constitute a single homogeneous group with respect to their price movements. In so far as they are to be used for further production, or to undergo further manufacture, they resemble raw materials in relation to the bidding of competing manufacturers, and their prices, therefore, are characterized by relatively wide oscillations. In so far as the demand for them is for the purpose of final consumption, purely business forces have less weight, and their prices are more stable. Related to this argument is that which has already been presented, the increasing stability of prices as the stable elements of wages and overhead charges bulk larger in commodity costs. So, again, the sample price quotations from which an index of wholesale prices is to be constructed must include prices representative of producers' and consumers' goods, of goods in the intermediate as well as the final stages of manufacture.

Other important divisions of the price system exist. The behavior of the prices of capital equipment differs from that of prices of goods intended for human consumption. The

[1] It should not be inferred from this that there is no relation between agricultural production and the prices of agricultural products, and general business conditions. The immediate price relation is frequently one of contradictory movements, and cycles in agricultural production are not synchronous with business cycles. But conditions in these two fields of economic activity are mutually related in many ways.

[2] Cf. "The Making and Using of Index Numbers," 47.

prices of durable goods differ in their fluctuations from the prices of perishable goods. Goods imported into a given country and goods exported from that country are usually subject to the play of different forces. A representative index number of wholesale prices should be based upon price quotations drawn from all commodity groups marked by distinctive modes of behavior, with weight given to each in proportion to the relative importance in trade of the commodities in that category.

Price Comparisons over Time

In the opening pages of this chapter the fact was noted that the degree of dispersion found in frequency distributions of price relatives depended upon the length of time covered in price comparisons. Hence, on statistical grounds, there is justification for the conclusion that the accuracy of well-constructed price indices is high for measurements extending over a short interval, and becomes progressively lower as the range of the time comparison increases. This conclusion is supported by other considerations.

In Laspeyres' formula,

$$I = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}$$

the price factor alone varies, as between numerator and denominator. The constant weighting factor, q_0, is assumed to define quantities entering into trade in an unchanging system of income distribution, living standards, consumption habits, etc. This system, for which Sir George Knibbs has used the term "regimen," is taken to be common to the two periods compared. If it is constant, and if the q's which define its quantitative attributes are unchanged, then we may expect to measure with accuracy the one factor which does change — commodity prices. The condition we have here assumed is the orthodox one of *ceteris paribus*, the condition that factors other than the one subject to study remain constant.

In fact, of course, the regimen does not remain fixed.

Changes in tastes and in consumption habits occur; changes in types of goods used as capital equipment take place; incomes shift, and the flow of goods is altered by changes in the distribution of buying power among consuming groups; the very price changes that we seek to measure bring alterations in the demand for given types of goods, and in the quantities produced. Of no small moment in the total situation are the changes that occur in the quality of goods that continue to pass by the same trade names. The automobile of 1938 is the same commodity, by name, as the automobile of 1910, but to the average consumer the later model represents quite a different bundle of utilities. Similarly, steel, textiles, locomotives, even the staple articles of diet have undergone important quality changes. A comparison of price levels in 1910 and 1938 that depends for its accuracy on the assumption that all elements of economic life except prices have remained constant is suspect, indeed.

Our difficulties are not removed if we take as the standard of reference the regimen of the second of the two periods compared. This is done in Paasche's formula,

$$I = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}.$$

The system of consumption standards and all that goes with it may be of modern vintage in this case, but the differences between the regimens of the two periods compared is just as wide. We have not held constant non-price factors, and our measurement of price changes loses in accuracy, as a result.

The method exemplified by the Ideal formula, that of employing weighting factors drawn from both periods, represents one attempt at the solution of this problem, but it is far from perfect. The use of quantities drawn from the two regimens does not create a common regimen, the indispensable condition of full accuracy in such comparisons.

The practical procedure in the face of this difficulty is to restrict our comparisons, if high accuracy is required, to

periods not widely separated in time. Consumption habits, living standards and technical production methods will be not widely dissimilar in two such periods, and hence the number of identical commodities common to the two periods will be large. Under these conditions considerable confidence may be placed in index numbers measuring average price changes. Comparison of price levels over longer periods may be desired, and may be justified, but the margin of error in the measurements may be expected to increase as the time span extends. Formal precision in weighting and in the selection of acceptable formulas will not provide an escape from the unavoidable difficulties arising out of alterations in the basic conditions of economic life. Real continuity of indices covering a stretch of years is possible only on the basis of a persisting common regimen.

These considerations support the claims of an index of the chain type, which involves the measurement of price changes between successive periods not far apart in time. Bruce D. Mudgett has advocated this procedure. The comparison of price levels in two periods, close together in time and with similar regimens, will be accurate, if such an index as the Ideal be employed. The elements of such a chain may then be linked together, in attempting to measure price changes between non-consecutive periods. If the regimens of the non-consecutive periods differ materially, the accuracy of the comparison will probably not be high. But it is reasonable to believe that better results will be secured by bridging the intervening years in the manner proposed than by constructing a single far-flung index based only upon the widely dissimilar regimens of two periods far removed in time.

The Wholesale Price Index of the United States Bureau of Labor Statistics

The authoritative index of wholesale prices in the United States is that compiled by the United States Bureau of Labor

Statistics. This index was first constructed in 1902, for the period beginning with 1890. It was continued until 1913 as an unweighted average of relative prices, the base of each relative being the average price of the given commodity during the ten-year period 1890–1899. Various revisions of procedure have been made since 1913. As it stands at present the index for any given period (week, month, or year) is a weighted aggregate of actual prices, the aggregate being expressed, to facilitate comparison, as a relative with 1926 as the base.

The index now includes 813 price series. (A single commodity may be represented by several quotations, the prices for different grades or in different markets being given. Thus for raw cotton there are three quotations, Middling, New Orleans; Middling Upland, New York; and Middling Upland, Galveston.) In the derivation of the aggregate for any date each price quotation is multiplied by a given weight, known as a "quantity weight" or a "multiplier." This same weight is applied to the price quotation for the base period. The cross products thus obtained for the base period and the date in question are values of a stated quantity of goods; they differ only in respect of the price factor. The following tabulation illustrates the method as applied to cotton:

Commodity	*Average price, November, 1937 (per pound)*	*Quantity weight (pounds)*	*Average price, November, 1937 × quantity weight*
	p_k	q_k	$p_k q_k$
Cotton, Middling, New Orleans	$.080	1,399,496,000	$111,959,680
Cotton, Middling Upland, N. Y.	.080	77,750,000	6,220,000
Cotton, Middling Upland, Galveston	.077	6,297,729,000	484,925,133

When this process is carried out for the entire 813 price series included, the sum of the values in the last column gives the index number for the given period, in this case

November, 1937. As published, this sum is expressed as a relative, the aggregate in 1926 representing 100.[1] The formula for the index measuring the level of wholesale prices at time "1," with reference to the base level at time "0" is, thus,

$$\frac{P_1}{P_0} = \frac{\Sigma p_1 q_h}{\Sigma p_0 q_h}$$

where q_h represents the constant multipliers. The method of construction renders it possible to shift the base to any desired year or month, changing the given relatives to percentages on the new base.

This index number, therefore, is based upon the cost at wholesale of a bill of goods. The bill of goods remaining the same, the total cost changes as the prices of the various commodities change, and the index measures the effect of these changing individual prices upon the total cost.

It is essential, of course, that the quantity used as multiplier for each series of price quotations truly represent the relative importance of the commodity in question. The multipliers employed are approximations to the quantities actually marketed. Changes are made from time to time in these quantities, the revisions being applied, of course, to the base period aggregates as well as to the figures for later periods. In addition, when it is necessary to substitute one price series for a related one that has been discontinued or has lost significance, minor modifications are made in the multipliers so as to maintain comparability between the aggregates for periods preceding and periods following the date of substitution.[2]

The Bureau of Labor Statistics publishes index numbers of wholesale prices for 10 major and 45 minor commodity

[1] This base is, at the date of writing, twelve years removed in time. Adoption of a 1935–1937 base period is now being considered.

[2] The method of adjustment is explained in an article, "Revised Method of Calculation of the B. L. S. Wholesale Price Index," by Jesse M. Cutts and Samuel J. Dennis, *Journal of the American Statistical Association*, December, 1937, 663–674.

groups, as well as a general index for all commodities. The major groups include farm products, foods, hides and leather products, textile products, fuel and lighting, metals and metal products, building materials, chemicals and drugs, house furnishing goods, and miscellaneous commodities. The constituent elements of the index are also classified into raw materials, semi-manufactured articles and finished products, and measurements of price changes for these groups are computed. The National Bureau of Economic Research has constructed index numbers for various other categories of commodities, utilizing the quotations of the Bureau of Labor Statistics. These classes include raw and processed goods, durable and non-durable goods, producers' goods and consumers' goods, goods destined for use in capital equipment and goods destined for human consumption, foods and non-foods, and crops, animal products, minerals, and forest products.[1] The availability of index numbers for various significant classes of goods makes it possible to trace price changes with more precision, and to interpret them more accurately, than when dependence is placed upon a single all-embracing index. For the elements of the price system are marked by wide diversity in their behavior over both long and short periods of time.

Other Price Index Numbers

The measurement of price changes by the use of index numbers has not been confined to wholesale prices. Many variations of this device have been utilized in measuring price movements in other fields. It will be useful at this point briefly to indicate the character of some of these variations.[2]

[1] See *Prices in Recession and Recovery*, N. Y., National Bureau of Economic Research, 1936, 492–540.

[2] Detailed information concerning the character and content of a wide variety of index numbers, price and other, will be found in *An Index to Business Indices*, Donald H. Davenport and Frances V. Scott, Chicago, Business Publications, Inc., 1937.

INDEX NUMBERS OF RETAIL PRICES

An index of retail food prices is published currently by the United States Bureau of Labor Statistics. The general methods employed are similar to those already explained in connection with the index of wholesale prices computed by that agency, with such differences as inevitably result from the nature of the material.

Actual retail selling prices of 84 articles of food are secured biweekly from dealers in 51 representative cities throughout the United States. In weighting the quotations on foods of a single type (fresh vegetables, for example) in a given city, account is taken of the quantities of such foods consumed by an average wage-earner's family in that city or, for some regions, in the district in which that city lies. In obtaining weights consumption by food groups is considered, rather than by specific commodities, since the commodities actually priced must be taken to represent similar commodities for which no prices are collected.

The combination for a single city (or geographical area) of food prices thus weighted yields an index for that region. The food cost index for the United States is computed from the aggregates for the 51 cities, each weighted according to the population of the area which the city is taken to represent. Thus the weights entering the final index of retail food prices for the country as a whole represent quantities consumed by the average wage-earner's family, and the population assumed to be affected by each series of quoted prices. The base of the index numbers, as published, is the average of the three-year period 1923–1925.

The indices of retail food prices, together with index numbers of the prices of electricity and coal, at retail, are published in the *Monthly Labor Review*.

The difficulties inherent in the problem of measuring wholesale price movements have been discussed at some length. The construction of index numbers of retail prices of the type just described presents even greater difficulties.

All the theoretical problems arising in the former case are to be solved and, in addition, the practical difficulties of securing suitable weights, accurate price figures, and comparable quotations are intensified. Because of the lack of commodity standardization, and because of variations in business practice and local customs, the latter difficulty is particularly acute. For these reasons no index of retail prices at present published can be accepted with the confidence with which the best indices of wholesale prices may be received.

INDEX NUMBERS OF THE COST OF LIVING

If these problems are acute in constructing an index of retail prices they are doubly hard to solve in measuring such an entity as the cost of living. When food prices, rents, retail clothing prices, cost of fuel and light, retail furniture prices, and the prices of the other miscellaneous items which are included in the budget of the average family are to be averaged, and an index number constructed to measure variations in the cost of these items, numerous statistical difficulties must be overcome. Theoretical questions concerning the most suitable methods of averaging and weighting present themselves, but more important are the practical problems involved in the collection of accurate and comprehensive prices and weighting data.

Two index numbers of the cost of living are currently compiled in the United States, one by the Bureau of Labor Statistics, one by the National Industrial Conference Board of New York. The former appears in the *Monthly Labor Review*, the latter in periodic publications of the Conference Board. In each case the chief items of domestic expenditure are weighted in accordance with their relative importance in household budgets, and the combined results expressed as relative numbers. These are given on the 1913 and 1923–1925 base by the Bureau of Labor Statistics, on the 1923 base by the Conference Board.[1]

[1] For a general discussion of the problem, with details of the Conference

INDEX NUMBERS OF PRICE AND BUYING POWER OF FARM PRODUCTS

A set of useful index numbers relating to the prices received by and the prices paid by farmers is compiled by the United States Department of Agriculture. The first of these is based upon the prices at the farm, as of the middle of each month, of 34 major farm products and 13 commercial truck crops. The weights employed are the average quantities marketed by farmers during the period 1924–1929. Farmers and agricultural economists have need of such a specialized index, because the wholesale prices of farm products in the great exchanges or in large cities are often poor representatives of the prices actually received by farmers.

The index of prices paid by farmers is compiled quarterly (in March, June, September, and December). The constituent quotations are retail prices paid by farmers for commodities used in family maintenance and in production. Weights are estimated quantities bought by farmers. The base of the index of farm prices, as published, is the average of the five pre-war years from August, 1909 to July, 1914; that of the index of prices paid, 1910–1914. Measurements for subgroups are given, in both cases.

These two index numbers are used in the derivation of an index of the purchasing power of farm products. The computation of the purchasing power index may be illustrated with reference to the figures for 1936. In that year the index of prices of farm products was 114. The index of prices paid by farmers was 124. That is, the farmer was receiving 14 per cent more, on the average, for a unit of product than in 1909–1914, but the average price paid by him for a unit of goods purchased was 24 per cent higher than in the base period. Therefore the purchasing power of an average unit of farm products was 8 per cent less than in 1909–1914 ($114 \div 124 = .92$).

Board procedure, see *Cost of Living in the United States, 1914–1936*, M. Ada Beney, New York, National Industrial Conference Board, 1936.

These three index numbers, for selected years, are given in Table 57.

TABLE 57

Index Numbers of Farm Prices, Prices Paid by Farmers, and the Buying Power of Farm Products [1]

(1) *Year*	(2) *Prices received by farmers*	(3) *Prices paid by farmers*	(4) *Average per unit purchasing power of farm products* (2) ÷ (3)
1910–1914	100 *	100	100
1918	202	176	115
1920	211	201	105
1921	125	152	82
1925	156	157	99
1929	146	153	95
1932	65	107	61
1933	70	109	64
1934	90	123	73
1935	108	125	86
1936	114	124	92
1937	121	131	93

* Aug., 1909–July, 1914 = 100.

These are significant measurements, yielding valuable information concerning the buying and selling relations that vitally affect one important group of producers. The development of similar measurements for other groups will add materially to our understanding of the changes that shifting market relations entail, in the economy at large. Yet the limitations of these index numbers should not be overlooked. The measurement of prices paid by farmers and, correspondingly, the measurement of the purchasing power of farm products, are subject to the difficulties referred to in the discussion of retail prices and living costs. Under existing conditions the margin of error in all such measurements is fairly wide. The error is the greater, too, the longer the time span covered by the quotations. In the present case, goods bought by farmers have undergone greater changes

[1] Source: *The Agricultural Situation*, U. S. Bureau of Agricultural Economics.

in quality than have the fairly standardized staples that the farmer sells. Here, as in all price comparisons over time, greater confidence must attach to short-period comparisons than to those spanning several decades.

REFERENCES

Bowley, A. L., *Elements of Statistics*, Chap. 9.

Chaddock, R. E., *Principles and Methods of Statistics*, Chap. 10.

Crum, W. L. and Patton, A. C., *An Introduction to the Methods of Statistics*, Chaps. 8, 19.

Davies, G. R. and Crowder, W. F., *Methods of Statistical Analysis in the Social Sciences*, Chap. 5.

Davies, G. R. and Yoder, Dale, *Business Statistics*, Chap. 3.

Day, E. E., *Statistical Analysis*, Chaps. 21–23.

Ferger, Wirth F., "Distinctive Concepts of Price and Purchasing-Power Index Numbers," *Journal of the American Statistical Association*, Vol. 31, No. 194, June, 1936, 258–272.

Fisher, Irving, *The Making of Index Numbers.*

Flux, A. W., "The Measurement of Price Changes," *Journal of the Royal Statistical Society*, March, 1921, 167–215.

Haberler, G., *Der Sinn der Indexzahlen.*

Kelley, T. L., *Statistical Method*, Chap. 13.

Keynes, J. M., *A Treatise on Money*, Chaps. 4–8.

King, W. I., *Index Numbers Elucidated.*

Knibbs, Sir George, "The Nature of an Unequivocal Price-Index and Quantity-Index," *Journal of the American Statistical Association*, March, June, 1924.

Mills, F. C., *The Behavior of Prices*, Chaps. 1, 3.
Prices in Recession and Recovery.

Mitchell, W. C., "The Making and Using of Index Numbers," Part I, *Bulletin 284*, U. S. Bureau of Labor Statistics, Oct. 1921.

Persons, W. M., "Fisher's Formula for Index Numbers," *Review of Economic Statistics*, Prel. Vol. III, 103–113.

Rhodes, E. C., *Elementary Statistical Methods*, Chap. 10.

Staehle, H., "A Development of the Economic Theory of Price Index Numbers," *Review of Economic Studies*, June, 1935. "International Comparisons of Food Costs," Appendix I (In: *Studies and Reports of the International Labour Office*, Series N, no. 20).

Walsh, C. M., *The Measurement of General Exchange Value.*
The Problem of Estimation.

Waugh, A. E., *Elements of Statistical Method*, Chap. 8.

CHAPTER VII

THE ANALYSIS OF TIME SERIES: MEASUREMENT OF TREND

The preceding sections have dealt primarily with frequency series and with problems arising in the attempt to organize and describe such series. We are now concerned with data in the study of which the essential problem is the analysis of chronological variations. Such series are of major importance in the field of economic statistics, for most of the data of economics and business are variables in time — as bank clearings, steel production, volume of sales, etc. This dominating importance of series in time is not found in any other field of statistical research, and the development of methods of analysis appropriate to time series has come, accordingly, only within recent years with the wider adoption of statistical methods in the field of economics.

Problems connected with time series arise both in the ordinary routine of internal administration and in the analysis of general economic conditions. Sales, purchases, profits on the one hand, stock prices, interest rates, business failures on the other, are variables which fluctuate with the passage of time. In the analysis of such series it is generally desired that the rate and character of growth be determined, and that periodic and accidental fluctuations be isolated for study. The sales manager wishes to know how the volume of sales is faring, when and why it fluctuates and how it compares with volume of production. The economist desires to trace the trend of prices, and to scrutinize minutely the upward and downward movements of the price level. The making of business plans on even a small scale, as

well as the most elaborate schemes of economic forecasting, must rest upon such study of past trends and fluctuations, and upon comparison of the movements of related series in time. Scientific study of the business cycle is only possible through the application of such methods. Our present task is the development of methods appropriate to the analysis of series in time.

The Preliminary Organization of Time Series

The data of time series usually require less preliminary organization than do statistical data which are to be reduced to the form of a frequency distribution. The source, primary or secondary, from which the figures are taken usually presents them in shape for analysis. Certain precautions should be observed, however.

The dates to which the figures apply should be clearly understood and definitely stated. Monthly data may be as of the first of each month (as for the stock price index of the New York Stock Exchange), averages for each month (as for the Bureau of Labor Statistics' price index), or totals for each month (as in the case of figures on cotton consumption). They may be cumulative monthly figures, each item representing the total for the year to date, as in the case of certain coal production data. If average figures are given for a month or year it is important to know how the average has been secured.

Again, it is essential that in any time series there be strict comparability between data for different periods. Any attempt to analyze a series that is not homogeneous must be misleading and futile. Yet such series are not infrequently published. Commodity production or consumption figures published by trade associations and by governmental agencies are sometimes based upon returns from a varying number of reporting concerns. A series of price quotations may lack comparability as between different dates because of changes in the unit or grade to

which the quotations apply, or because quotations are drawn from different markets. Changes in census classifications may result in lack of comparability of census data. A change in a salesman's territory may alter his returns materially. It is stated that the character of the obligations represented by the United States Steel Corporation's figures for "unfilled orders" has varied from time to time. Records relating to the physical output of a given commodity in different periods may be rendered inaccurate by changes in quality or design. These are examples of faults that may be found in time series, rendering analysis futile. Strict testing is essential before a series be accepted as accurate and homogeneous.

GRAPHIC REPRESENTATION OF TIME SERIES

Normally the first step to be taken in visualizing a series in time and in preparing for further analysis consists of plotting the data. The trend and general characteristics of a series may be most readily apprehended through graphic representation. The data may be plotted on ordinary arithmetic or on semi-logarithmic paper. The advantages of the latter types for certain purposes have already been explained. The choice in a given case will depend upon the nature of the data and the object of the study. If interest lies in the absolute amount of fluctuations in sales, prices, pig iron production or whatever may be in process of analysis, or in the comparison of absolute differences between series, the ordinary rectilinear chart is to be employed. If percentage variations and the comparison of relative fluctuations are matters of interest, the semi-logarithmic representation is preferable. In general, if one is accustomed to the interpretation of this latter type of chart, its use is advisable. A clearer, less-distorted presentation of relations and a more significant comparison of series are generally secured when economic data having time as one variable are plotted on paper with a logarithmic ruling on one axis.

For some purposes the process of studying series in time will have been completed when the data are thus plotted. The general trend may be roughly determined from the chart. The existence of seasonal and other periodic variations may be ascertained. Rough comparisons of trends and fluctuations may be made. All the knowledge thus secured, it should be noted, will be non-quantitative in character, and the comparisons will be tentative and approximate. Even so, such charts enable trends and relations to be much more clearly visualized than do the raw figures, and for some purposes the knowledge thus secured is sufficient, though it lacks precision and accuracy. For other purposes more exact measurement and more refined analysis are required. Certain appropriate methods may be described.

Forces Affecting Series in Time

The general object in the analysis of a time series is the isolation of the effects of one or more of the forces affecting the given series. This may be desired in order that the past behavior of the single series may be understood, in order that the future behavior of the series may be predicted, or in order that two or more series may be compared. It is not in any case possible to isolate these effects of individual forces with absolute accuracy, and in some cases it is impossible even to approximate such a result. But given figures covering a sufficiently long period, the effects of various influences upon the behavior of a given series may usually be measured with some degree of accuracy.

What are these forces that affect series of data in time? The forces in any given case may be unique, affecting only the given series, but in general the various influences acting upon such series may be placed in a limited number of categories.

SECULAR TREND

In the first place, most series of economic statistics exhibit definite trends. Such a trend may be constant in direction, may change direction at a constant rate, or may even be characterized by abrupt shifts in direction or rate that reflect the introduction of novel elements. Thus the volume of production or sales of a business house over a period of years usually shows a fairly regular growth. The same is true of population, the production of basic minerals, the number of motor vehicles registered, etc. In some cases the rate of growth may be a negative one, as is true of interest rates in the United States over the last half century. The concept of *secular trend* (i.e., trend over a long period of time) covers both positive and negative changes of this type.

In the analysis of a time series the trend value at any date is taken to be the "normal" value at that date. This conception of a normal value which may be used as a base or point of reference in judging the effects of all forces other than the growth factor, is fundamental in economic analysis. "No other method," says Carl Snyder, "enables us so quickly to set economic events in their just perspective." We should note, however, that such a normal value is essentially an empirical construction. While useful for purposes of reference, and as one of a series of measurements reflecting secular movements in a given series, it should not be assumed to possess any special normative significance.

The fact should be emphasized that by secular trend is meant the smooth, regular, long-term movement of a statistical series. Frequent and sudden changes either in absolute amounts or in rates of increase or decrease are quite inconsistent with the idea of secular trend. It is true that there may be occasional changes due to the interjection of a new element or the withdrawal of an old factor. But the breaking up into numerous sections of the period covered

by a time series, and the determination of trend for each of these minor periods, does violence to the very concept of secular change.

It does not follow from this discussion that a definite rising or declining trend exists for all time series. Many series, such as barometric readings at a certain point, merely fluctuate about a constant level that does not change with the passage of time.

PERIODIC FLUCTUATIONS

If the plotted representation of a time series be studied, the long-term trend may be discerned in the general upward or downward drift, but may not be precisely determined by inspection because of the existence of numerous fluctuations, superimposed upon the trend. These fluctuations may be regular or irregular, violent or mild, simple or complex. The value of the variable at any given date may be thought of as the net resultant of the interaction of the secular trend and the various forces that tend to modify the persistent secular movements of a given series. (It may be, in fact, that for many series the trend is the resultant of the interplay of a variety of conflicting forces, rather than an underlying movement upon which the periodic and other fluctuations are superimposed. In the present discussion no attempt is made to define the organic relations that may exist among the forces affecting a series in time.) These latter forces may be of several types.

Seasonal variations are found in most series of economic statistics for which quarterly, monthly, or weekly values are obtainable. Consumption and production of commodities, interest rates, bank clearings, railroad freight traffic, and many other types of data are marked by seasonal swings repeated with minor variations year after year. These, in so far as they exist at all, are definitely periodic in character, with a constant twelve-month period. Less markedly periodic, but nevertheless characterized by a

considerable degree of regularity, are the *cyclical fluctuations* that are found in series affected by forces connected with economic or business cycles. Prices, wages, the volume of industrial production, trading on the Stock Exchange, and most series relating to the activities of individual business units are affected by the swings of business through alternating periods of depression and prosperity. While the length of such periods may vary, the general sequence of change has been in the past sufficiently regular to render these cyclical movements capable of study.

RANDOM FLUCTUATIONS

Entangled with these more or less regular movements are the effects of random, accidental, and irregular fluctuations — catastrophic events such as the San Francisco earthquake, wars, floods, fires, and countless minor events equally fortuitous though less violent in the resulting disruptions. These events influence the value of a variable at a given date, modifying the effects of long-term movements and of seasonal and cyclical factors. The observed value at any time is the resultant of the play of all these forces.

The analysis of series in time involves the isolation of the effects of these various forces, so far as this is possible. A problem may call for the study of but one factor, or it may require the complete breaking up of given values. When annual data are used the seasonal element will not enter, of course. The explanation of methods begins with a consideration of problems involving only this type of data.

The Measurement of Secular Trend

As an example of the type of material in connection with which these problems arise, the figures in Table 58 on page 232 may be taken. The values are given in thousands of millions in order to simplify the calculations.

As has been pointed out, the figure for any year, as the

TABLE 58

New York Clearing House Transactions, 1875–1936

(In thousands of millions)

1875	25.1	1891	34.1	1907	95.3	1923	214.6
1876	21 6	1892	36.3	1908	73.6	1924	235.5
1877	23.3	1893	34.4	1909	99.3	1925	276.9
1878	22.5	1894	24.2	1910	102.6	1926	293.4
1879	25.2	1895	28.3	1911	92.4	1927	307.2
1880	37.2	1896	29.4	1912	96.7	1928	368.9
1881	48.6	1897	31.3	1913	98.1	1929	456.9
1882	46.6	1898	39.9	1914	89.8	1930	399.5
1883	40.3	1899	57.4	1915	90.8	1931	287.7
1884	34.1	1900	52.0	1916	147.2	1932	177.3
1885	25.3	1901	77.0	1917	181.5	1933	154.6
1886	33.4	1902	74.8	1918	174.5	1934	162.7
1887	34.9	1903	70.8	1919	214.7	1935	174.4
1888	30.9	1904	59.7	1920	252.3	1936	186.5
1889	34.8	1905	91.9	1921	204.1		
1890	37.7	1906	103.8	1922	213.3		

value of $162.7 thousands of millions for 1934, is the net resultant of the many forces that we have classified under the headings of secular trend, cyclical variations, and random or accidental fluctuations. Our first problem is to measure the secular trend.

In Fig. 54 the data of New York bank clearings during the period 1875–1936, inclusive, have been plotted. A definite trend is apparent, together with well marked and more or less regular deviations from that trend. Several methods are available for arriving at approximations to this trend. By employing moving averages an attempt may be made to eliminate passing fluctuations and to arrive at values that define the influence of the steadily operating growth factor. If we assume that a definite functional relationship prevails (empirically at least) between the time factor and the other variable, an approximation to the trend may be secured by fitting an appropriate curve to the plotted data. Smoothing the data by hand gives somewhat the same result, the curve being frankly approxi-

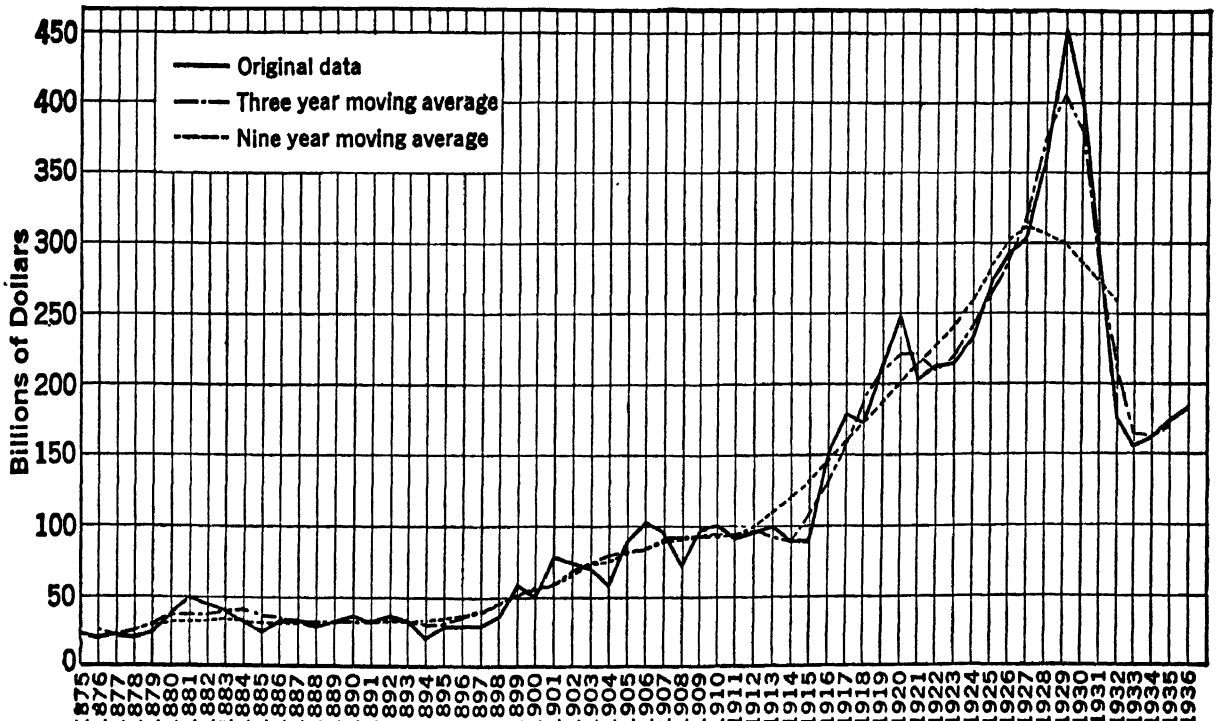


FIG. 54. — New York Clearing House Transactions, 1875–1936, with Moving Averages

mative and empirical in character. In certain studies it has been found possible to use one statistical series as base or trend line for another series of homogeneous data.

Moving Averages

When a trend is to be determined by the method of moving averages, the average value for a number of years (or months, or weeks) is secured, and this average is taken as the normal or trend value for the unit of time falling

Table 59

New York Clearing House Transactions, 1912–1936, *and* 3-, 5-, 7-, *and* 9-*Year Moving Averages*

(In thousands of millions)

Year	*Original data*	*Three-year moving av.*	*Five-year moving av.*	*Seven-year moving av.*	*Nine-year moving av.*
1912	$ 96.7				
1913	98.1	$ 94.87			
1914	89.8	92.90	$104.52		
1915	90.8	109.27	121.48	$125.51	
1916	147.2	139.83	136.76	142.37	$149.51
1917	181.5	167.73	161.74	164.40	161.44
1918	174.5	190.23	194.04	180.73	174.24
1919	214.7	213.83	205.42	198.23	188.11
1920	252.3	223.70	211.78	207.86	204.19
1921	204.1	223.23	219.80	215.57	218.60
1922	213.3	210.67	223.96	230.20	231.03
1923	214.6	221.13	228.88	241.44	245.78
1924	235.5	242.33	246.74	249.29	262.91
1925	276.9	268.60	265.52	272.83	285.64
1926	293.4	292.50	296.38	307.63	307.36
1927	307.2	323.17	340.66	334.04	315.62
1928	368.9	377.67	365.18	341.50	311.48
1929	456.9	408.43	364.04	327.27	302.49
1930	399.5	381.37	338.06	307.44	289.80
1931	287.7	288.17	295.20	286.80	276.58
1932	177.3	206.53	236.36	259.01	263.17
1933	154.6	164.87	191.34	220.39	
1934	162.7	163 90	171.10		
1935	174.4	174.53			
1936	[1]86.5				

at the middle of the period covered in the calculation of the average. Table 59 shows the results secured when three-, five-, seven-, and nine-year moving averages are thus computed for New York clearings for the period 1912–1936.

The three-year moving average for 1916 is the average of the figures for 1915–16–17, the five-year figure for 1916 is the average of the years 1914–15–16–17–18. The other averages are computed in the same way. In each case the average is centered for the period included; that is, the average is taken to represent normal as of the middle of the given period. The employment of an odd number of years simplifies this centering process, though it is not essential that the number be odd. With an even number of years, the figure may be centered by taking a two-year moving average of the first moving average. The three- and nine-year moving averages for the entire period are plotted, with the original data, in Fig. 54.

It is obvious that the effect of the averaging is to give a smoother curve, lessening the influence of the fluctuations that pull the annual figures away from the general trend. The longer the period included in securing each average, the smoother is the curve secured, though there are other factors to consider in deciding upon the length of the period. Certain of these factors may be noted.

CHARACTERISTICS OF MOVING AVERAGES

Given cyclical fluctuations about a uniform level or about a line ascending with a uniform slope, the length of the cycle and the magnitude of the fluctuations being constant, a moving average having a period equal to the period of the cycle (or to a multiple of that period) will give a straight line, a perfect representation of the trend. Under the same conditions a moving average having a period greater or less than the period of the cycle will give, not a straight line, but a new cycle having the same period as the original, but with fluctuations of less magnitude. The minima and

maxima of the cycles thus obtained will not necessarily coincide with the minima and maxima of the original cycles. In general, when such a new cycle is obtained the magnitude of the fluctuations will be less the longer the period on which the average is based.[1]

These propositions may be illustrated by the figures in Table 60, arbitrarily chosen. In the first example five figures have been selected which repeat themselves in sequence, fluctuating about a common level.

The moving averages in columns (2) and (3) represent

TABLE 60

Illustrating the Application of Moving Averages

(1) *Cyclical data*	(2) *Moving average of 5 items*	(3) *Moving average of 10 items (centered)*	(4) *Moving average of 3 items*	(5) *Moving average of 8 items (centered)*
2				
6			5⅓	
8	6⅕		8	
10	6⅕		7⅔	
5	6⅕		5⅔	6⅛
2	6⅕	6⅕	4⅓	6⅛
6	6⅕	6⅕	5⅓	6⅜
8	6⅕	6⅕	8	5¾
10	6⅕	6⅕	7⅔	5⅛
5	6⅕	6⅕	5⅔	6⅛
2	6⅕	6⅕	4⅓	6⅛
6	6⅕	6⅕	5⅓	6⅜
8	6⅕	6⅕	8	5¾
10	6⅕	6⅕	7⅔	5⅛
5	6⅕	6⅕	5⅔	6⅛
2	6⅕		4⅓	6⅛
6	6⅕		5⅓	
8	6⅕		8	
10			7⅔	
5				

(The items in columns (3) and (5) have been centered by means of a moving average of 2 items.)

[1] The decrease in the magnitude of the fluctuations is not regular, however, but cyclical.

the data with the cycles completely removed. When the period of the average is not equal to the period of the cycle, or to a multiple of that period, the cycle is not removed, as is apparent from the figures in columns (4) and (5).

The conclusions suggested above hold when the cyclical fluctuations take place about any straight line. In Table 61 the foregoing data have been employed but with a constant increment of 3. This is equivalent to superimposing the same cycles upon a line with a slope of + 3.

TABLE 61

Illustrating the Application of Moving Averages to a Series with Linear Trend

(1) *Cyclical data*	(2) *Moving average of 5 items*	(3) *Moving average of 10 items (centered)*	(4) *Moving average of 3 items*	(5) *Moving average of 8 items (centered)*
2				
9			8⅓	
14	12⅕		14	
19	15⅕		16⅔	
17	18⅕		17⅔	18⅛
17	21⅕	21⅕	19⅓	21 1/16
24	24⅕	24⅕	23⅓	24⅛
29	27⅕	27⅕	29	26¾
34	30⅕	30⅕	31⅔	29 11/16
32	33⅕	33⅕	32⅔	33⅛
32	36⅕	36⅕	34⅓	36 13/16
39	39⅕	39⅕	38⅓	39⅜
44	42⅕	42⅕	44	41¾
49	45⅕	45⅕	46⅔	44 11/16
47	48⅕	48⅕	47⅔	48⅛
47	51⅕		49⅓	51 13/16
54	54⅕		53⅓	
59	57⅕		59	
64			61⅔	
62				

(The items in columns (3) and (5) have been centered by means of a moving average of 2 items.)

The trend values, with the effect of the cycles completely removed, are secured by taking moving averages equal in

period to the cycle or to a multiple of that period. The cycle persists, with the same period but with diminished amplitude, when the average is based upon a period not equal to that of the cycle, as is clear from the figures in columns (4) and (5).

When these ideally simple conditions of constant period and amplitude do not exist, the moving average becomes more ambiguous and its interpretation less simple. If the period of the cycle varies, the selection of a period for the moving average is more difficult. In general, a period equal to or greater than the average length of the cycle is to be selected. An average having a shorter period will give a line that is marked by pronounced cycles, these cycles being reduced as the period covered in the calculation of the average increases.

When the amplitude of the cycle varies, the period being constant, a moving average with a period equal to the length of the cycle will give a line of trend marked by minor cycles. The amplitude of these secondary cycles will be a minimum when the period of the average is equal to the period of the cycle (or to a multiple of that period). When these last two irregularities are combined, and the data are characterized by cycles of varying amplitude and of varying length, the moving average giving the most effective representation of the trend is that which has a period equal to the average length of the cycle, or to a multiple of that length.

A new factor enters when the trend departs from linearity. If the underlying trend of a series is concave upward, a moving average will always exceed the actual trend value; if the reverse is true, and the trend is convex upward, a moving average will always be less than the actual trend value.

These conditions are depicted in the following examples. The figures in Table 62 give the values secured when a cycle of constant period and amplitude, as in col. (3), is

superimposed upon a line of trend that is concave upward, i.e., increasing at a constantly increasing rate. If the moving average could completely eliminate the effects of the cycle, the values secured from the average would be equal to the average value of the five items in each cycle (6.2) plus the values of the function $y = x^2$, given in col. (2).

TABLE 62

Illustrating the Application of Moving Averages to a Non-Linear Series

(Increasing rate)

(1) x	(2) x^2	(3) *Cyclical data*	(4) *Col. (2) plus col. (3)*	(5) *Moving average of 5 items*	(6) *True trend values* $(x^2 + 6.2)$
0	0	2	2		
1	1	6	7		
2	4	8	12	12.2	10.2
3	9	10	19	17.2	15.2
4	16	5	21	24.2	22.2
5	25	2	27	33.2	31 2
6	36	6	42	44.2	42.2
7	49	8	57	57.2	55.2
8	64	10	74	72 2	70.2
9	81	5	86	89.2	87.2
10	100	2	102	108.2	106.2
11	121	6	127	129.2	127 2
12	144	8	152	152 2	150.2
13	169	10	179	177.2	175.2
14	196	5	201	204.2	202.2
15	225	2	227	233.2	231.2
16	256	6	262	264.2	262.2
17	289	8	297	297.2	295.2
18	324	10	334		
19	361	5	366		

The values of the moving average are, in this case, in excess of the true trend values, a form of distortion that will always occur with a series of this type.

In Table 63 are shown the results of superimposing the same cyclical values upon a line of trend that is convex upward, i.e., increasing at a constantly decreasing rate.

In this case, a perfect method of eliminating the cycles would give results equal to the average value of the five items (6.2) plus the values of the function $y = \sqrt{x}$.

In this case the moving average values are consistently too low. The discrepancy is greatest for the lower values of x, as the decrease in the rate of growth is most marked for these values.

TABLE 63

Illustrating the Application of Moving Averages to a Non-Linear Series

(Decreasing rate)

(1) x	(2) $\sqrt{x}$	(3) *Cyclical data*	(4) *Col. (2) plus col. (3)*	(5) *Moving average of 5 items*	(6) *True trend values* ($\sqrt{x} + 6.2$)
0	0	2	2.00		
1	1.00	6	7.00		
2	1.41	8	9.41	7.428	7.61
3	1.73	10	11.73	7.876	7.93
4	2 00	5	7.00	8.166	8.20
5	2.24	2	4.24	8.414	8.44
6	2.45	6	8.45	8.634	8.65
7	2.65	8	10.65	8.834	8.85
8	2.83	10	12.83	9.018	9.03
9	3.00	5	8.00	9.192	9.20
10	3.16	2	5.16	9.354	9.36
11	3.32	6	9.32	9.510	9.52
12	3.46	8	11.46	9.658	9.66
13	3.61	10	13.61	9.800	9.81
14	3.74	5	8.74	9.936	9.94
15	3.87	2	5.87	10.068	10.07
16	4.00	6	10.00	10.194	10.20
17	4.12	8	12.12	10.318	10.32
18	4.24	10	14.24		
19	4.36	5	9.36		

Considerations previously reviewed have indicated that a moving average should, in general, be based upon a period at least equal to the period of the cycle, and preferably equal to some higher multiple of that period when the

data are at all irregular. The longer the period covered, the greater the stability of the average. But when the underlying trend departs materially from the linear form, following a curve bending upward or downward, the error involved in the use of any moving average increases as the period of the average increases. If a moving average is used in such a case to measure the trend, the period of the average should be the shortest which will serve to average out the cycles; equal, that is, to the average length of one cycle.

In practice, however, these various conditions are found in complicated combinations. The fact that cycles vary in amplitude and length calls for a moving average based upon a fairly long period. The fact that the trend of the data is usually non-linear calls for a short period average to lessen the upward or downward distortion. A consideration of some importance in practical work is that a moving average can never be brought up to date. The lag is less, of course, the shorter the period covered by the average. The selection of a period in a given case must rest upon a study of the actual data with these various considerations in mind.

It has been assumed in the preceding discussion that the purpose of the moving average is the representation of secular trend. The moving average may be used, also, in smoothing data for the purpose of eliminating random fluctuations. For this purpose a moving average based upon a period shorter than the average length of the cycle should be selected.

We may return now to the problem relating to New York bank clearings. A study of the lines marked out by the different moving averages in Fig. 54 reveals significant differences between them. The three-year average follows the graph of the original data most closely, as would be expected. The nine-year average marks out the smoothest line of trend, but, on the other hand, departs most widely

from the data. This is particularly noticeable from 1893 to 1898, from 1911 to 1915, from 1921 to 1926, and from 1927 to 1931. It is due to the pronounced changes in the rate of growth of the series during these periods. Except for these distortions the general trend seems to be most accurately represented by the nine-year average.

In determining the relative merits of the different moving averages we are aided by a knowledge of the course of business during the period covered. The volume of New York bank clearings is a sensitive index of general business conditions, responding immediately to changes in speculative and industrial activity. Major and minor business cycles are reflected in this series. Knowing the number of cycles through which business has passed during the period 1875–1936, we may determine which of the moving averages serves best as a standard from which to measure cyclical deviations. In this case we are practically working backward from a known result, a method not always available.

If we take as a starting point in each cycle the year in which revival began, after recession, the following cycles in general business activity may be distinguished:[1]

1871–1879	1908–1912
1879–1885	1912–1914
1885–1888	1915–1919
1888–1891	1919–1921
1891–1894	1921–1924
1894–1897	1924–1927
1897–1900	1927–1933
1901–1904	1933–
1904–1908	

The cycles marked out by the three-year moving average are too numerous to enumerate. In fact, the deviations from this average are primarily accidental and minor

[1] These dates are based upon the chronology of American business cycles developed by Wesley C. Mitchell; cf. "Production during the American Business Cycle of 1927–1933," by Wesley C. Mitchell and Arthur F. Burns, *Bulletin 61*, National Bureau of Economic Research, November 9, 1936. It should be noted that the chronology is based upon monthly data, whereas the Clearing House data cited in the text are annual figures.

fluctuations and should not be classed as cycles. Deviations from the five-, seven-, and nine-year averages mark out the following cycles:

Cycles of deviations from five-year moving averages	*Cycles of deviations from seven-year moving averages*	*Cycles of deviations from nine-year moving averages*
1879–1885	1879–1885	1879–1885
1885–1888	1885–1888	1885–1888
1888–1891	1888–1894	1888–1897
1891–1897	1894–1900	1897–1900
1897–1900	1900–1904	1900–1904
1900–1904	1904–1908	1904–1908
1904–1908	1908–1911	1908–1915
1908–1911	1911–1915	1915–1923
1911–1915	1915–1918	1923–
1915–1918	1918–1923	
1918–1924	1923–1927	
1924–1927	1927–1932	
1927–1932	1932–	
1932–		

Some of the differences between the series of cycles thus determined and the reference cycles distinguished by Mitchell are doubtless due to the distinctive behavior of New York clearings. Other differences reflect the peculiarities of moving averages. Deviations from the five-year averages between 1879 and 1927 show one more cycle than we find in the series based on seven-year averages, four more cycles than are shown by the nine-year averages. And yet the deviations from five-year averages fail to show the cycles of 1894–1897 and of 1921–1924. The nine-year averages reveal only eight cycles between 1879 and 1927, as against Mitchell's fourteen reference cycles. Mitchell was working, of course, with monthly data which are more sensitive than annual data to cyclical forces. Moreover, he was dealing with relatively short movements, some of which appear as only minor fluctuations in general business activity.

If interest attaches to the shorter swings of business, to cycles with average durations of three or four years,

a moving average of relatively short period should be used. A five-year average is appropriate. Averages of longer period define trend movements more faithfully, but may fail to reveal fluctuations properly classified as business cycles. We should refer, however, to recent attempts to establish the reality of long cycles, of nine, eleven, or as many as thirty years in average duration. In the study of such cycles moving averages of corresponding periods would be employed.

In general, the moving average has the prime advantage of flexibility. The representation of secular trend by mathematical curves frequently involves the breaking up of a period into two or three subdivisions, and the fitting of separate curves to each. This results from changing conditions and sharply changing rates of growth or decline. Where such changes occur the moving average has the merit of flexible adaptation to the new conditions and is often a more effective measure of secular trend than curves fitted with great labor.

Simple and weighted moving averages, in varying combinations, have wide uses in the analysis of economic time series. An illuminating discussion of these uses, and of the procedures appropriate to different purposes, is to be found in *The Smoothing of Time Series*, by Frederick R. Macaulay.[1]

Representation of Secular Trend by Mathematical Curves

For many types of data the secular trend may be represented by a mathematical curve rather than by a line based upon a moving average. Thus, if the growth (or decline) is by constant absolute increments (or decrements) a straight line will serve as an exact representation of the trend. Or the growth may be by constant percentages, as in the case of capital increase, when a principal sum increases in accordance with the compound interest law.

[1] National Bureau of Economic Research, New York. 1931.

A curve of a definite mathematical form furnishes the best representation of this trend. In many series of economic statistics the data seem to conform to definite laws of growth, or decline, and where this is the case the task of analysis, interpretation, and projection is materially assisted by securing a mathematical expression for the underlying law. In practically all cases, of course, there are departures from this law, deviations above and below the line of secular trend. These deviations, however, do not destroy the value of an equation that describes the underlying law of development.

There is one fundamental difference between the moving average as a measure of trend and such mathematical curves. The former implies no definite "law" to which the data are assumed to conform. It is based upon the data as given; if the general trend changes, the moving average follows the new trend. It is a flexible measure of trend, adapting itself to changing conditions, purporting to be nothing more than an empirical approximation to the drift of the series. Mathematical curves fitted to economic series are, in fact, nothing more than empirical approximations also, but in a somewhat different sense. They assume a "law" of change underlying the variations, accidental and otherwise, which show upon the surface of the data. It is an empirical law which is assumed, it is true, but nevertheless there is postulated a uniform and consistent trend capable of mathematical expression. If such an assumption is to have any validity it is essential that the period during which the law is supposed to hold be homogeneous, that there be no material changes in the conditions affecting the series being studied. Thus an equation is secured for the trend of gold production, let us say. If a radical change should take place in methods of extraction the trend of gold production would change materially and the former equation would no longer apply. Data covering the period before and after such a change

would not be homogeneous, and a single equation for the trend during the whole period should not be secured.

In the practical approach to a problem involving the determination of secular trend the first task is the selection of the appropriate type of curve. This is perhaps the most difficult part of the work; certainly it is the part in which the element of personal judgment enters most directly. For there is no objective rule to follow, no fixed standard by which the most appropriate curve may be selected. Something more will be said on this subject after the characteristics of the chief types of curves and the methods of fitting them have been described. For the present it may be assumed that a curve similar to one of the types described in Chapter II, or to a related form, has been selected, and that we face the practical task of fitting it to the data.

FITTING A STRAIGHT LINE; THE METHOD OF LEAST SQUARES

If the data, when plotted, show a trend that can best be represented by a straight line the task of fitting is merely the determination of the constants in an equation of the form $y = a + bx$. The values of a and b which will give a line following most closely the trend of the data are to be obtained. A simple illustration may serve to demonstrate the various methods which may be employed. Nine points (1, 3; 2, 4; 3, 6; 4, 5; 5, 10; 6, 9; 7, 10; 8, 12; 9, 11) are plotted in Fig. 55. Our problem is the fitting of a straight line to these points.

By inspection approximate values of a and b may be determined. A thread may be stretched through the points in such a direction that it seems to follow the trend as closely as possible. The slope of the line thus laid out may be measured, the y-intercept determined, and the desired equation thus approximated. Obviously this is a loose and uncertain method, and the results obtained by different individuals may be expected to vary rather widely.

There is one and only one straight line that fits the plotted data most accurately. The constants for this line of best fit may be determined by the method of least squares.

The theory upon which the method of least squares is based need not be detailed at length here. The argument may be briefly presented: A number of observation values of a certain quantity are found, and it is desired to obtain the most probable value of the quantity which is being

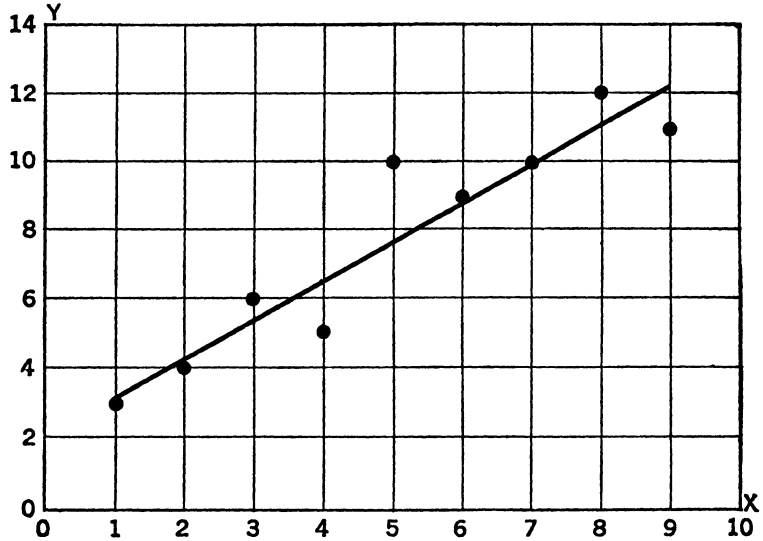


FIG. 55. — Illustrating the Fitting of a Straight Line to Nine Points

measured. It is capable of demonstration that the most probable value of the quantity is that value for which the sum of the squares of the residuals is a minimum. (The "residual" is a term for the difference between a given estimated value and an observation value.) This is true of the arithmetic mean of the observation values. Thus, if a given distance be measured by a number of individuals, with varying results, the most probable value is the arithmetic mean of the different measurements. The process of computing the mean involves the following steps, which are enumerated for the purpose of simplifying the later

explanation. We seek a result, a statement of the most probable value of the distance being measured, which will take the form:

$$M = \text{(a constant)}.$$

Let us say we have three approximations to this value:

$$\begin{aligned} M &= 5{,}672 \text{ feet} \\ M &= 5{,}671 \text{ feet} \\ M &= 5{,}676 \text{ feet} \\ \text{adding,} \quad 3M &= 17{,}019 \text{ feet.} \end{aligned}$$

Since there is but one unknown, M, it may be derived directly from this equation, and we have

$$M = 5{,}673 \text{ feet}.$$

This is the value for which the sum of the squares of the deviations is a minimum.

A similar problem arises when the relation between two variables is being measured. Our goal in this case is the equation that correctly describes this relationship. We have secured, however, varying results which do not agree precisely as to the constants in the equation of relationship. In other words, our plotted points do not all lie on the same line. What are the most probable values of the constants in the required equation? The answer is analogous to that given when a single quantity was being measured. We seek the constants which, when the resulting equation is plotted, will give a line from which the deviations of the separate points, when squared and totaled, will be a minimum. Assuming that each pair of measurements gives an approximation to the true relationship between the variables, we wish to find the most probable relationship, and this is given by the line for which the sum of the squared deviations is a minimum.[1]

We have, in the present example, nine pairs of values for x and y. Substituting these values in the generalized form

[1] Cf. Appendix A for a more detailed discussion of the method of least squares, together with a description of certain checks upon the calculations.

of the linear equation, $y = a + bx$, we secure the following observation equations:

$$\begin{aligned} 3 &= a + 1b \\ 4 &= a + 2b \\ 6 &= a + 3b \\ 5 &= a + 4b \\ 10 &= a + 5b \\ 9 &= a + 6b \\ 10 &= a + 7b \\ 12 &= a + 8b \\ 11 &= a + 9b. \end{aligned}$$

Any two of these equations could be solved as simultaneous equations, and values of a and b secured. But these values would not satisfy the remaining equations. Our problem is to combine the nine observation equations so as to secure two *normal equations*, which, when solved simultaneously, will give the most probable values of a and b. The first of these normal equations is secured by multiplying each of the observation equations by the coefficient of the first unknown (a) in that equation, and adding the equations obtained in this way. Since the coefficient of a in the present case is 1 throughout, the nine observation equations are unchanged by the process of multiplication. The second of the normal equations is secured by multiplying each of the observation equations by the coefficient of the second unknown (b) in that equation, and adding the equations obtained. Thus the first equation is multiplied throughout by 1, the second by 2, and so on. The process of securing the two normal equations is illustrated in Table 64 on page 250.

The two normal equations are

$$\begin{aligned} 70 &= 9a + 45b \\ 418 &= 45a + 285b. \end{aligned}$$

It remains to solve these equations for a and b. By multiplying the first equation by 5 and subtracting it from the second, a may be eliminated; a value of $\frac{68}{60}$, or 1.133, is

TABLE 64

Derivation of Normal Equations from Observation Equations

$3 = a + 1b$	$3 = 1a + 1b$
$4 = a + 2b$	$8 = 2a + 4b$
$6 = a + 3b$	$18 = 3a + 9b$
$5 = a + 4b$	$20 = 4a + 16b$
$10 = a + 5b$	$50 = 5a + 25b$
$9 = a + 6b$	$54 = 6a + 36b$
$10 = a + 7b$	$70 = 7a + 49b$
$12 = a + 8b$	$96 = 8a + 64b$
$11 = a + 9b$	$99 = 9a + 81b$
$70 = 9a + 45b$	$418 = 45a + 285b$

found for b. Substituting this value in either of the equations, a value of 2.111 is secured for a. The equation to the best fitting straight line is, therefore,

$$y = 2.111 + 1.133x.$$

In the actual application of the method it is not necessary to write out and total the equations, as is done above. We need only insert the proper values in the two equations,[1]

$$\Sigma(y) = na + b\Sigma(x)$$
$$\Sigma(xy) = a\Sigma(x) + b\Sigma(x^2).$$

The symbols employed have the following meanings:

$\Sigma(y)$: the sum of the values of y.
$\Sigma(x)$: the sum of the values of x.
$\Sigma(xy)$: the sum of the products of the paired x's and y's.
$\Sigma(x^2)$: the sum of the squares of the values of x.
n: the number of pairs of values; the number of points plotted.

The work of computation is facilitated by a tabular arrangement similar to that shown in Table 65.

The two desired normal equations are secured by substituting these five values in the type equations given above. It will be noted that the results are identical with those obtained from the observation equations.

[1] General rules for the formation of normal equations are given in Appendix A.

TABLE 65

Computation of Values Required in Fitting a Straight Line

x	y	xy	x^2	
1	3	3	1	$n = 9$
2	4	8	4	$\Sigma(x) = 45$
3	6	18	9	$\Sigma(y) = 70$
4	5	20	16	$\Sigma(x^2) = 285$
5	10	50	25	$\Sigma(xy) = 418$
6	9	54	36	
7	10	70	49	
8	12	96	64	
9	11	99	81	
45	70	418	285	

When the equation to the best fitting straight line has been obtained the values of y corresponding to given values of x may be computed and compared with the observed values. Table 66 presents the results secured:

TABLE 66

Comparison of Observed and Computed Values of a Variable Quantity[1]

x	y (*observed*)	y (*computed*)	d	d^2	xd
1	3	$3.2\frac{4}{5}$	$-\ .2\frac{4}{5}$	.0597	$-\ .2\frac{4}{5}$
2	4	$4.3\frac{1}{5}$	$-\ .3\frac{1}{5}$	.1427	$-\ .7\frac{1}{5}$
3	6	$5.5\frac{1}{5}$	$+\ .4\frac{1}{5}$	.2390	$+\ 1.4\frac{3}{5}$
4	5	$6.6\frac{4}{5}$	$-\ 1.6\frac{4}{5}$	2.7041	$-\ 6.5\frac{1}{5}$
5	10	$7.7\frac{1}{5}$	$+\ 2.2\frac{1}{5}$	4.9381	$+\ 11.1\frac{1}{5}$
6	9	$8.9\frac{1}{5}$	$+\ .0\frac{1}{5}$	.0079	$+\ .5\frac{2}{5}$
7	10	$10.0\frac{2}{5}$	$-\ .0\frac{2}{5}$	.0020	$-\ .3\frac{1}{5}$
8	12	$11.1\frac{2}{5}$	$+\ .8\frac{2}{5}$	.6760	$+\ 6.5\frac{1}{5}$
9	11	$12.3\frac{1}{5}$	$-\ 1.3\frac{1}{5}$	1.7190	$-\ 11.8$
			0.0	10.4885	0.0

The sum of the deviations of the plotted points from the line is zero. The sum of the deviations when each is multiplied by the corresponding value of x is also zero. The accuracy of the actual calculations involved in fitting may

[1] The common fractions are retained in certain columns in order that the sum of the deviations may be exactly zero.

be tested in this way. The sum of the squares of the deviations, 10.4885, is a minimum. Any change in the value of a or b would give a line from which the sum of the squared deviations would exceed 10.4885.

FITTING A STRAIGHT LINE; SPECIAL CASES

The simultaneous solution of the two normal equations will give, in any case, the most probable values of a and b. The processes of calculation may be simplified in certain special cases, not infrequently encountered in handling economic data. If the x's are consecutive numbers, as they always are when an unbroken time series is plotted, the origin may be taken at the median value. When the number of observations is odd this will be the middle item, of course. The value of $\Sigma(x)$ will then be zero, and the normal equations become

$$\Sigma(y) = na$$
$$\Sigma(xy) = b\Sigma(x^2).$$

Thus if a time series extends, by years, from 1901 to 1937, the origin may be taken at 1919, the value of x corresponding to 1918 being -1, to 1920, $+1$, and so on. The solution for values of a and b is rendered much easier when the data may be disposed in this way. When there is an even number of years the same process is possible, time (the x-variable) being measured in units of one half year.

Again, when the values of x are consecutive positive numbers starting at zero, the values of $\Sigma(x)$ and of $\Sigma(x^2)$ may be easily determined. The sum of the first n natural numbers is equal to $\frac{n(n+1)}{2}$. Thus the sum of the numbers from 1 to 5 is $\frac{5(5+1)}{2}$, or 15. This term may replace $\Sigma(x)$ in the normal equations. Similarly, the sum of the squares of the first n natural numbers is equal to $\frac{2n^3+3n^2+n}{6}$. Thus the sum of the squares of the numbers from 1 to 5

is equal to $\frac{250 + 75 + 5}{6} = 55$. This expression may replace $\Sigma(x^2)$ in the normal equations, and we have

$$\Sigma(y) = na + b\left(\frac{n(n+1)}{2}\right).$$

$$\Sigma(xy) = a\left(\frac{n(n+1)}{2}\right) + b\left(\frac{2n^3 + 3n^2 + n}{6}\right).$$

It is sometimes easier to work from equations in this form than in the form first given. The data for time series may be handled in this way, the years being numbered consecutively, beginning with 1.

FITTING A CURVE OF THE POWER SERIES

The discussion above has been confined to the case of linear trend. Such a type of curve frequently gives an excellent fit, but in many cases it fails accurately to fit the data. This difficulty is sometimes overcome in practice by breaking a series into segments and fitting a separate line to the data for each of these periods. Where there is an actual break in the series, the period as a whole lacking homogeneity, this practice may be justified, but when the period is essentially homogeneous the whole concept of secular trend is violated by this process of subdividing and fitting separate lines. In many cases where a straight line will not fit, a curve of the power series may represent the trend accurately. The general process of fitting such a curve may be briefly described.

The generalized form of the equation of the type desired is $y = a + bx + cx^2 + dx^3 + \ldots$. An equation of this form does not, of course, represent a curve of the parabolic type, but in ordinary usage that term is applied to the potential series. If carried to the second power of x it is called a second degree parabola; if to the third power, a third degree parabola, etc. For ordinary purposes such a curve should not be carried beyond the second or third power of x.

If carried to the second power there are three unknowns, and three normal equations must be solved simultaneously in securing the required values.

The procedure is similar to that outlined for the linear case. Each observation equation is multiplied by the coefficient of the first unknown in that equation, and the resulting equations are totaled to give the first normal equation. The process is repeated for the two other unknowns, and the three normal equations thus obtained are solved for a, b, and c. The results are the most probable values of these three constants. The following are the general forms which the three normal equations take:

$$\Sigma(y) = na + b\Sigma(x) + c\Sigma(x^2).$$
$$\Sigma(xy) = a\Sigma(x) + b\Sigma(x^2) + c\Sigma(x^3).$$
$$\Sigma(x^2y) = a\Sigma(x^2) + b\Sigma(x^3) + c\Sigma(x^4).$$

As an example of the process, the calculations involved in fitting a second degree parabola to the points 1, 2; 2, 6; 3, 7; 4, 8; 5, 10; 6, 11; 7, 11; 8, 10; 9, 9 may be outlined. It is of the greatest practical importance in curve fitting, as in all extensive calculations, that the work be laid out and carried on in a definite and systematic fashion, with each step definitely related to the preceding and succeeding operations. Checks should be introduced wherever possible, as mathematical errors creep into even the most careful work. A tabular arrangement is generally advisable, each operation being revealed and each set of results clearly presented. The data in the present problem may be arranged as in Table 67.

When the x's are consecutive integers beginning with 1, as in the present case, the values of $\Sigma(x)$, $\Sigma(x^2)$, $\Sigma(x^3)$, and $\Sigma(x^4)$ may be secured from prepared tables.[1]

[1] Cf. Table XXVIII, Pearson, *Tables for Statisticians and Biometricians*. Cambridge University Press; Tables D and E, Mills and Davenport, *Manual of Problems and Tables in Statistics*, New York, Henry Holt and Co. Values to the third power are given in Appendix Table IX of the present volume.

TABLE 67

Computation of Values Required in Fitting a Second Degree Parabola

x	y	xy	x^2	x^2y	
1	2	2	1	2	$n = 9$
2	6	12	4	24	$\Sigma(x) = 45$
3	7	21	9	63	$\Sigma(x^2) = 285$
4	8	32	16	128	$\Sigma(x^3) = 2{,}025$
5	10	50	25	250	$\Sigma(x^4) = 15{,}333$
6	11	66	36	396	$\Sigma(y) = 74$
7	11	77	49	539	$\Sigma(xy) = 421$
8	10	80	64	640	$\Sigma(x^2y) = 2{,}771$
9	9	81	81	729	
45	74	421	285	2,771	

Substituting these values in the equations given above, the following normal equations are secured:

$$74 = 9a + 45b + 285c.$$
$$421 = 45a + 285b + 2{,}025c.$$
$$2{,}771 = 285a + 2{,}025b + 15{,}333c.$$

When these equations are solved simultaneously the following values are secured for the three constants:

$$a = -.929.$$
$$b = +3.523.$$
$$c = -.267.$$

The equation of the desired curve is

$$y = -.929 + 3.523x - .267x^2.$$

This curve and the nine given points are plotted in Fig. 56 on page 256.

If the values of x are consecutive, as in the present example, the work of computation is lightened if the mid-value is taken as origin. In this case $\Sigma(x)$ and $\Sigma(x^3)$ are equal to zero, and the normal equations become

$$\Sigma y = na + c\Sigma(x^2).$$
$$\Sigma(xy) = b\Sigma(x^2).$$
$$\Sigma(x^2y) = a\Sigma(x^2) + c\Sigma(x^4).$$

When a third degree parabola of the form $y = a + bx + cx^2 + dx^3$ is to be fitted to data, four constants must be determined, and four normal equations are necessary. These are of the following form:

$$\Sigma(y) = na + b\Sigma(x) + c\Sigma(x^2) + d\Sigma(x^3).$$
$$\Sigma(xy) = a\Sigma(x) + b\Sigma(x^2) + c\Sigma(x^3) + d\Sigma(x^4).$$
$$\Sigma(x^2y) = a\Sigma(x^2) + b\Sigma(x^3) + c\Sigma(x^4) + d\Sigma(x^5).$$
$$\Sigma(x^3y) = a\Sigma(x^3) + b\Sigma(x^4) + c\Sigma(x^5) + d\Sigma(x^6).$$

The solution for four or more constants involves a considerable amount of arithmetical calculation, and there is

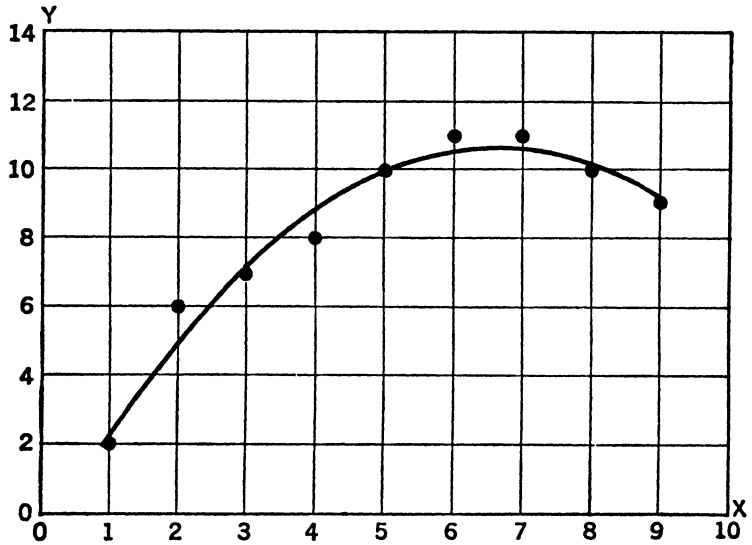


FIG. 56. — Illustrating the Fitting of a Second Degree Curve to Nine Points

some question as to the advisability of representing secular trend by equations of this type. With a sufficient number of constants a curve may be fitted which will follow every variation in the data, but such a curve could hardly be taken to represent the long time trend.[1] Minor departures from

[1] Regarding the employment of potential series of the type indicated for representing empirical curves, Steinmetz states that their use is justified:
1. If the successive coefficients a, b, c . . . decrease in value so rapidly that

a simple and uniform trend, linear or otherwise, are to be expected with economic data, but, if a real trend exists, extreme departures from a fairly simple form are rare. If such departures are due to pronounced changes in conditions no single line of trend is likely to be satisfactory, and it is advisable to break the period into parts, with a separate line of trend for each part. "Empirical curves," says Steinmetz, "can be represented by a single equation only when the physical conditions remain constant within the range of the observations." Though this statement relates to the fitting of curves to data from the physical sciences, it applies equally well to economic data.

Determination of the Secular Trend of a Business Series

Fitting a Straight Line

The procedure of fitting certain types of curves to simple data has been illustrated in the preceding sections. Before proceeding to a discussion of slightly different forms, it will be helpful to examine concrete examples of trend determination. We first determine the secular trend of a series defining the number of concerns in business in the United States, during the period from 1899 to 1914.[1] The observations are given in Table 68, together with the values required for the fitting of a straight line to the data, and the derived trend values. The values of x represent the time factor, while the values of y are the corresponding numbers of business concerns. Only the entries in columns

---

(*Footnote 1 continued from page 256.*)
within the range of observation the higher terms become rapidly smaller and appear as mere secondary terms.

2. If the successive coefficients follow a definite law, indicating a convergent series which represents some other function, as an exponential, trigonometric, etc.

3. If all the coefficients are very small, with the exception of a few of them, and only the latter ones thus need to be considered. Cf. C. P. Steinmetz, *Engineering Mathematics*, New York, McGraw-Hill, 1917, 214–215.

[1] Data compiled by Dun and Bradstreet.

(2) to (5), it should be noted, are employed in the fitting process.

TABLE 68

Number of Concerns in Business in the United States, 1899–1914

Computation of values required in fitting line of trend

(1) *Year*	(2) x	(3) y *No. of concerns in business, in thousands*	(4) xy	(5) x^2	(6) y_c *Trend values (linear) of no. of concerns in business, in thousands*
1899	1	1,148	1,148	1	1,152
1900	2	1,174	2,348	4	1,184
1901	3	1,219	3,657	9	1,217
1902	4	1,253	5,012	16	1,250
1903	5	1,281	6,405	25	1,283
1904	6	1,320	7,920	36	1,316
1905	7	1,357	9,499	49	1,349
1906	8	1,393	11,144	64	1,382
1907	9	1,418	12,762	81	1,415
1908	10	1,448	14,480	100	1,448
1909	11	1,486	16,346	121	1,481
1910	12	1,515	18,180	144	1,513
1911	13	1,525	19,825	169	1,546
1912	14	1,564	21,896	196	1,579
1913	15	1,617	24,255	225	1,612
1914	16	1,655	26,480	256	1,645
Totals	136	22,373	201,357	1,496	

$N = 16$ $\quad \Sigma(y) = 22{,}373$

$\Sigma(x) = 136$ $\quad \Sigma(xy) = 201{,}357$

$\Sigma(x^2) = 1{,}496$

The equations to be solved in determining the required constants are of the form

$$\Sigma(y) = Na + b\Sigma(x)$$
$$\Sigma(xy) = a\Sigma(x) + b\Sigma(x^2).$$

Inserting the given values in the formulas, we have

$$22{,}373 = 16a + 136b$$
$$201{,}357 = 136a + 1{,}496b$$

from which

$$a = 1{,}118.65$$
$$b = 32.90.$$

The equation to the line of best fit is therefore

$$y = 1{,}118.65 + 32.90x$$

with origin at 1898.

The trend values derived from this equation appear in column (6) of Table 68. The original data and line of trend are plotted in Fig. 57. The fit for the period covered

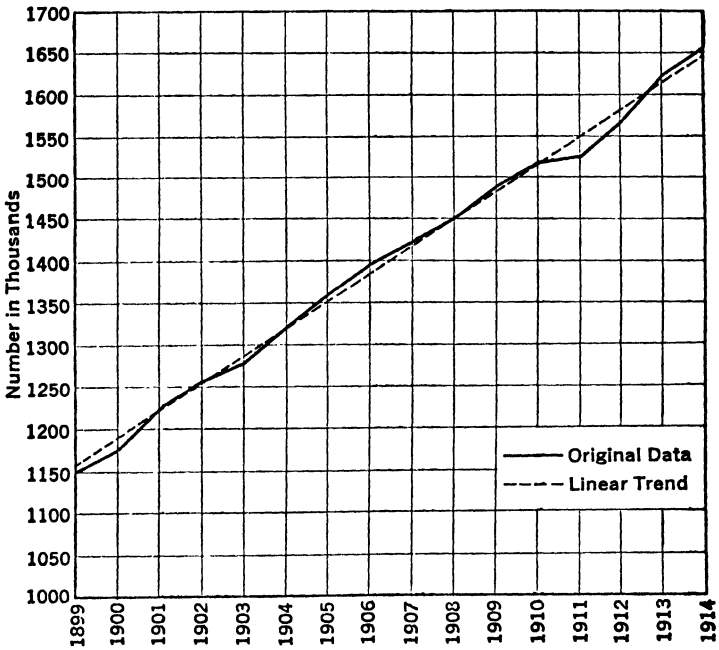


FIG. 57. — Number of Concerns in Business in the United States, 1899–1914, with Line of Trend

is good. The number of concerns in business in the United States during the sixteen years before the World War is well defined by the straight line we have secured.

FITTING A POWER CURVE OF THE SECOND DEGREE

The record of commercial failures in the United States over the last forty years provides an example of a series following a definitely non-linear trend. Data for the period 1897–1933 are presented in Table 69, with accompanying computations.

TABLE 69

Commercial Failures in the United States, 1897–1933 [1]

Computation of values required in fitting line of trend

(1) *Year*	(2) x	(3) y (*No. of failures*)	(4) xy	(5) x^2y
1897	− 18	13,351	− 240,318	4,325,724
1898	− 17	12,186	− 207,162	3,521,754
1899	− 16	9,337	− 149,392	2,390,272
1900	− 15	10,774	− 161,610	2,424,150
1901	− 14	11,002	− 154,028	2,156,392
1902	− 13	11,615	− 150,995	1,962,935
1903	− 12	12,069	− 144,828	1,737,936
1904	− 11	12,199	− 134,189	1,476,079
1905	− 10	11,520	− 115,200	1,152,000
1906	− 9	10,682	− 96,138	865,242
1907	− 8	11,725	− 93,800	750,400
1908	− 7	15,690	− 109,830	768,810
1909	− 6	12,924	− 77,544	465,264
1910	− 5	12,652	− 63,260	316,300
1911	− 4	13,441	− 53,764	215,056
1912	− 3	15,452	− 46,356	139,068
1913	− 2	16,037	− 32,074	64,148
1914	− 1	18,280	− 18,280	18,280
1915	0	22,156	0	0
1916	1	16,993	16,993	16,993
1917	2	13,855	27,710	55,420
1918	3	9,982	29,946	89,838
1919	4	6,451	25,804	103,216
1920	5	8,881	44,405	222,025
1921	6	19,652	117,912	707,472
1922	7	23,676	165,732	1,160,124
1923	8	18,718	149,744	1,197,952
1924	9	20,615	185,535	1,669,815

[1] Dun and Bradstreet.

TABLE 69—*Continued*

Commercial Failures in the United States, 1897–1933

(1) Year	(2) x	(3) y	(4) xy	(5) x^2y
1925	10	21,214	212,140	2,121,400
1926	11	21,773	239,503	2,634,533
1927	12	23,146	277,752	3,333,024
1928	13	23,842	309,946	4,029,298
1929	14	22,909	320,726	4,490,164
1930	15	26,355	395,325	5,929,875
1931	16	28,285	452,560	7,240,960
1932	17	31,822	540,974	9,196,558
1933	18	19,626	353,268	6,358,824
Totals		610,887	+ 1,817,207	75,307,301

$$N = 37 \qquad \Sigma(x^4) = 864{,}690$$
$$\Sigma(x) = 0 \qquad \Sigma(y) = 610{,}887$$
$$\Sigma(x^2) = 4{,}218 \qquad \Sigma(xy) = 1{,}817{,}207$$
$$\Sigma(x^3) = 0 \qquad \Sigma(x^2y) = 75{,}307{,}301$$

The origin is taken at the middle year to facilitate the calculations. The values of $\Sigma(x^2)$ and $\Sigma(x^4)$ may be secured from prepared tables, or from the formulas cited on page 254.

The normal equations to be solved in fitting a second degree parabola, with the origin at the middle year of the period covered, are of the form

$$\Sigma(y) = Na + c\Sigma(x^2)$$
$$\Sigma(xy) = b\Sigma(x^2)$$
$$\Sigma(x^2y) = a\Sigma(x^2) + c\Sigma(x^4).$$

Inserting the appropriate values, we have

$$610{,}887 = 37a + 4{,}218c$$
$$1{,}817{,}207 = 4{,}218b$$
$$75{,}307{,}301 = 4{,}218a + 864{,}690c.$$

Solving for the constants,

$$a = 14{,}827.6$$
$$b = 439.82$$
$$c = 14.762.$$

The required equation is

$$y = 14{,}827.6 + 430.82x + 14.762x^2$$

with origin at 1915.

The original observations and the line of secular trend

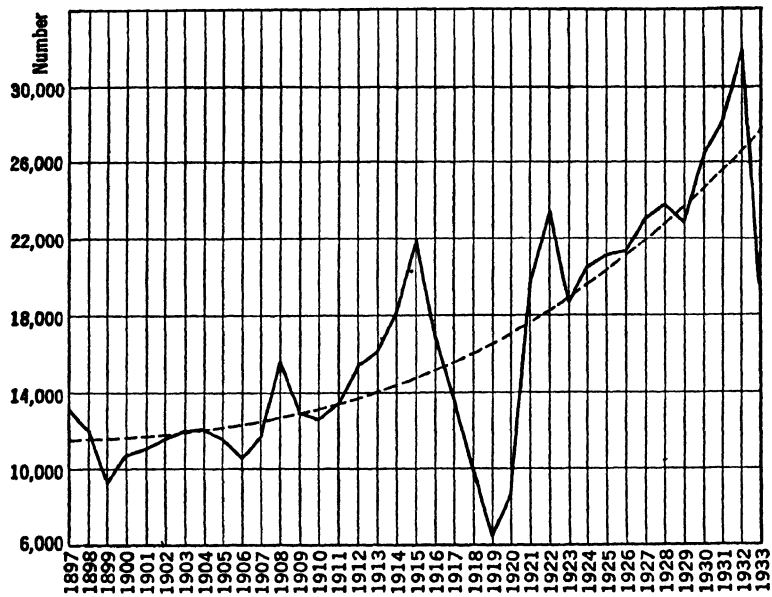


FIG. 58. — Commercial Failures in the United States, 1897–1933, with Line of Trend

are plotted in Figure 58. Observations, trend values and deviations from trend are given in Table 70.

Commercial failures reflect the major cycles in American business, but with movements that reverse those of most economic series. Failures are numerous in times of depression, fewer in prosperity. The reader who will compare the deviations from trend shown in Table 70 with the dates of reference cycles given on an earlier page will note the general agreement. The sharp fall in business failures from 1932 to 1933 reflected, of course, the special conditions prevailing in the latter year.

TABLE 70

Commercial Failures in the United States, 1897–1933

Actual Values, Trend Values, and Deviations from Trend

Year	*Number of commercial failures*	*Trend value, second degree parabola*	*Deviation of actual from trend value*
1897	13,351	11,855.73	+ 1,495.27
1898	12,186	11,769.88	+ 416.12
1899	9,337	11,713.55	− 2,376.55
1900	10,774	11,686.75	− 912.75
1901	11,002	11,689.47	− 687.47
1902	11,615	11,721.72	− 106.72
1903	12,069	11,783.49	+ 285.51
1904	12,199	11,874.78	+ 324.22
1905	11,520	11,995.60	− 475.60
1906	10,682	12,145.94	− 1,463.94
1907	11,725	12,325.81	− 600.81
1908	15,690	12,535.20	+ 3,154.80
1909	12,924	12,774.11	+ 159.79
1910	12,652	13,042.55	− 390.55
1911	13,441	13,340.51	+ 100.49
1912	15,452	13,668.00	+ 1,784.00
1913	16,037	14,025.01	+ 2,011.99
1914	18,280	14,411.54	+ 3,868.47
1915	22,156	14,827.60	+ 7,328.40
1916	16,993	15,273.18	+ 1,719.82
1917	13,855	15,748.29	− 1,893.29
1918	9,982	16,252.92	− 6,270.92
1919	6,451	16,787.07	− 10,336.07
1920	8,881	17,350.75	− 8,469.75
1921	19,652	17,943.95	+ 1,708.05
1922	23,676	18,566.68	+ 5,109.32
1923	18,718	19,218.93	− 500.93
1924	20,615	19,900.70	+ 714.30
1925	21,214	20,612.00	+ 602.00
1926	21,773	21,352.82	+ 420.18
1927	23,146	22,123.17	+ 1,022.83
1928	23,842	22,923.04	+ 918.96
1929	22,909	23,752.43	− 843.43
1930	26,355	24,611.35	+ 1,743.65
1931	28,285	25,499.79	+ 2,785.21
1932	31,822	26,417.76	+ 5,404.24
1933	19,626	27,365.25	− 7,739.25

The second degree curve employed to define the trend of commercial failures does so with reasonable accuracy over the period here covered. Extrapolation beyond those limits would be hazardous. Indeed, the changed conditions under which banking and many other types of business were conducted after 1933 may well break the continuity of the series, and generate a new long-term trend.

The Use of Logarithms in Curve Fitting

The family of curves described above represents a simple and very useful type. Perhaps of even greater general utility, in the analysis of time series, are curves of a semi-logarithmic type. The advantages of plotting many series of data on semi-logarithmic or "ratio" paper were explained in an earlier section. A fundamental virtue of this type of plotting is that it presents a true picture of *relative* variations, of *ratios* between magnitudes. Relations of this type are ordinarily of primary interest in the analysis of economic data, and it is logical that determination of trends should proceed on the same basis.

In doing so, we can make use of a group of curves of the same general form as those already described, the one difference being that log y takes the place of y throughout. That is, the straight line form is $\log y = a + bx$, while the general form for the potential series is $\log y = a + bx + cx^2 + dx^3 + \ldots$. The curves secured may be constructed on arithmetic paper, plotting the natural x's and the logarithms of the y's, or natural values of both x's and y's may be plotted on semi-logarithmic paper, the logarithmic scale extending along the y-axis. The latter is the simpler method.

To illustrate the procedure, the steps involved in fitting a curve of the type $\log y = a + bx$ will be shown. The trend of petroleum production in the United States from 1922 to 1929 is to be determined. The values needed in the normal equations are derived from Table 71.

TABLE 71

Petroleum Production in the United States, 1922–1929
(Computation of values required in fitting line of trend)

Year	x	y	$\log y$	$x \cdot \log y$
1922	1	557.5	2.74624	2.74624
1923	2	732.4	2.86475	5.72950
1924	3	713.9	2.85364	8.56092
1925	4	763.7	2.88292	11.53168
1926	5	770.9	2.88700	14.43500
1927	6	901.1	2.95477	17.72862
1928	7	901.5	2 95497	20.68479
1929	8	1,007.3	3.00316	24.02528
			23.14745	105.44203

$N = 8$ $\quad \Sigma(\log y) = 23.14745$
$\Sigma(x) = 36$ $\quad \Sigma(x \cdot \log y) = 105.44203$
$\Sigma(x^2) = 204$

The two normal equations to be solved are of the form

$$\Sigma(\log y) = Na + b\Sigma x$$
$$\Sigma(x \cdot \log y) = a\Sigma x + b\Sigma x^2.$$

Substituting the given values we have

$$23.14745 = 8a + 36b$$
$$105.44203 = 36a + 204b.$$

Solving for the constants,

$$a = 2.75645$$
$$b = .03044.$$

The equation to the desired curve is, therefore,

$$\log y = 2.75645 + .03044x$$

with origin at 1921.

In fitting this curve by the method of least squares, as is done above, we satisfy the condition that the sum of the squares of the *logarithmic* deviations shall be a minimum. That is, the deviations to which this condition relates are the differences between the logarithms of the observed values and the logarithms of the corresponding trend values.

This curve, it should be noted, is not the same as that from which the sum of the squares of the arithmetic (natural) deviations is a minimum.

The substitution in the above equation of the value of x representing any given year will enable the logarithm of the trend or normal value to be calculated. The trend value in natural numbers may then be determined. In Table 72 the normal value for each of the years covered is given, together with the percentage relation of actual to normal.

TABLE 72

Trend of Petroleum Production in the United States, 1922–1929, with Comparison of Actual and Trend Values

(Straight line trend determined from logarithms of production figures)

Year	x	y (*actual*) *Production (in millions of bbls.)*	$\log y_c$ *Log of trend*	y_c *(y, computed) Trend value (in millions of bbls.)*	*Percentage relation of actual to trend*
1922	1	557.5	2 78689	612.2	91.1
1923	2	732.4	2.81733	656.6	111.5
1924	3	713.9	2.84777	704.3	101.4
1925	4	763 7	2.87821	755.5	101 1
1926	5	770.9	2 90865	810 3	95.1
1927	6	901.1	2 93909	869.1	103.7
1928	7	901.5	2.96953	932.2	96.7
1929	8	1,007.3	2.99997	999.9	100.7

The points representing the actual production, together with the line of trend, are plotted in Fig. 59. The graph of the derived equation gives a good representation of the trend in the present instance.

An equation of this type, defining a linear trend in the logarithms of the dependent variable, has certain distinctive advantages. The reader will note that this is the logarithmic form of an equation to a compound interest curve (an exponential curve). This equation was given in Chapter II as

$$y = p(1 + r)^x$$

or

$$\log y = \log p + \log (1 + r)x.$$

In the example just given we have used the symbol a for $\log p$ and the symbol b for $\log (1 + r)$, but the equations are identical.

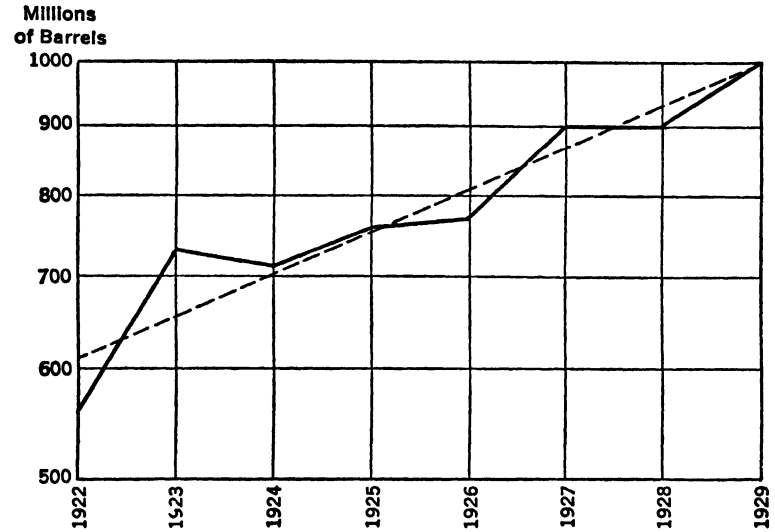


FIG. 59. — Production of Petroleum in the United States, 1922–1929, with Line Defining Average Rate of Growth

We may readily change to natural numbers the constants in the equation defining the trend of petroleum production from 1922 to 1929. We have

$$\log y = 2.75645 + .03044x$$

where 2.75645 is $\log p$ and .03044 is $\log (1 + r)$. The natural number corresponding to 2.75645 is 570.8; the natural number corresponding to .03044 is 1.0726. The trend of petroleum production in natural form is, therefore,

$$y = 570.8(1.0726)^x$$

with origin at 1921.

Subtracting 1 from the constant 1.0726 we secure .0726, which is r, the rate of increase of a series growing in accordance with the compound interest law. (If, on subtracting

1, we have a negative value, the growth is negative, of course.) This measure indicates that the production of crude petroleum increased at an average rate of 7.26 per cent a year between 1922 and 1929 (r being multiplied by 100 to place it on a percentage basis).

When the trend of a series in time may be defined by a straight line on ratio paper, and it is surprising how widely

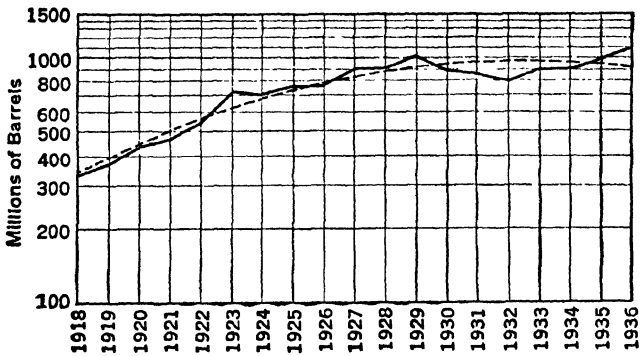


FIG. 60. — Production of Crude Petroleum in the United States, 1918–1936, with Line of Trend

applicable such a function is, the constant r is a highly useful measure. It defines the average annual rate of growth or decline of the series. It is, of course, an abstract measure and thus has the great merit of permitting comparison of the trends of series relating to widely different original units. The rate of growth of population, over a given period, may have been 1.4 per cent per year; the production of gasoline may have increased at a rate of 4.5 per cent, the production of automobiles at 4.2 per cent, the production of wheat at 1.1 per cent, total national income at 1.6 per cent, total national debt at 3.2 per cent. The trends of these series are immediately comparable, and conclusions concerning the direction and character of a nation's development may be drawn. This measure provides a valuable device for the study of social and economic change.[1]

[1] In any extensive application of this procedure time and labor may be

TABLE 73

Production of Crude Petroleum in the United States, 1918–1936
Comparison of Actual and Trend Values

(Trend values determined from second degree parabola fitted to logarithms of production figures)

Year	*y (actual) Production (in millions of bbls.)*	*y (computed) Trend value (in millions of bbls.)*	*Percentage relation of actual to trend*
1918	335.9	345.0	97.4
1919	378.4	395.5	95.7
1920	442 9	449.2	98.6
1921	472 2	505.2	93.5
1922	557.5	562 8	99.1
1923	732.4	620.8	118.0
1924	713.9	678.2	105.3
1925	763.7	733.8	104.1
1926	770.9	786.3	98.0
1927	901.1	834.3	108.0
1928	901.5	876.8	102.8
1929	1,007.3	912.4	110.4
1930	898.0	940.5	95.5
1931	850.3	960.0	88 6
1932	785.2	970.4	80.9
1933	905.7	971.5	93.2
1934	908.1	963.2	94.3
1935	996.6	945.7	105.4
1936	1,098.5	919.6	119.5

By the use of additional terms a function of the type just discussed may be modified, when dealing with a series marked by non-linear trends on ratio paper. For example, if the course of petroleum production be followed over a longer period, as is done in Fig. 60, it is obvious that the trend line secured for the period 1922–1929 is inappropriate. The addition of a third constant gives an equation of the type

$$\log y = a + bx + cx^2.$$

(*Footnote 1 continued from page 268.*)
saved by utilizing Glover's mean value table (cf. James W. Glover, *Tables of Applied Mathematics*, George Wahr, Ann Arbor, Michigan, 1923, 468ff.). By the use of this table the compound interest curve may be fitted directly to the natural numbers. All necessary computations are simply and quickly performed.

In fitting this to the data of petroleum production for the period 1918–1936, we may follow the exact procedure used when y was the dependent variable in a similar equation (see page 261), except that log y is used throughout, instead of y. For the required equation we have

$$\log y = 2.921331 + .023660x - .002107x^2$$

with origin at 1927. This is shown graphically in Fig. 60. Actual and trend values, in natural terms, are given in Table 73 on page 269.

Other Curve Types

The two families of curves described in the preceding sections meet most of the needs of the economic statistician. The trend in most time series may be described by curves of the power series, fitted either to natural numbers or to the logarithms of the data (that is, to the logarithms of the y values; time, the x-variable, is treated in terms of natural numbers in fitting both the above types of curves). These classes constitute flexible and widely applicable curve forms.[1] Attention may be called to several other curve types which have been applied less extensively to time series, but with favorable results in particular cases.

Curves of the ordinary parabolic type ($y = ax^b$) are not

[1] There are available for fitting higher degree curves of the power series methods that lessen the labor involved, particularly if curves of different degree are to be fitted to the same data. These methods, which reduce the fitting process to a series of simple adding machine operations, are appropriate to extended research projects. Their use is not advisable, however, unless work involving a considerable number of routine operations is contemplated. It is desirable that the student master the basic least squares procedures outlined in the preceding pages, utilizing other methods only in case extended computing tasks are undertaken.

For accounts of systematic methods suited to extensive calculations, see R. A. Fisher, *Statistical Methods for Research Workers*, Edinburgh, Oliver and Boyd, Sixth edition, 1936, 148–156; Max Sasuly, *Trend Analysis of Statistics: Theory and Technique*, Washington, Brookings Institute, 1934. The application of the method of orthogonal polynomials described by Fisher is admirably exemplified in James W. Angell, *The Behavior of Money*, New York, McGraw-Hill, 1936, 195–202.

generally applicable to economic data in the form of time series, as their use involves the treatment of the time variable as a geometric series. Such a curve, it will be recalled, becomes a straight line on double logarithmic paper. Yet if a curve of this form serves accurately to describe the trend of a given series, its use is justified, empirically.

Such curves may be fitted most readily by employing logarithms and using an equation of the linear type. The equation

$$y = ax^b$$

becomes, in logarithmic form,

$$\log y = \log a + b \log x.$$

The two normal equations needed in fitting such a curve are of the form

$$\Sigma(\log y) = n \log a + b\Sigma(\log x)$$
$$\Sigma(\log x \cdot \log y) = \log a\Sigma(\log x) + b\Sigma(\log x)^2.$$

By substituting the values computed from the data, these equations may be solved for log a and b, just as in fitting an ordinary straight line.[1]

The equation to the simple exponential curve may be written

$$y = ar^x.$$

(The r in this equation is the equivalent of $1 + r$, as given on p. 267.) This equation may be used to define the trend of a series increasing or decreasing in geometric progression. It has been observed that the trends of economic series frequently depart from such a geometric progression by constant magnitudes. By adding this magnitude, in a given case, to the original series (or subtracting it), a

[1] A useful table of the sums of the logarithms of the natural numbers from 1 to 100 is included as an appendix to *Medical Biometry and Statistics*, by Raymond Pearl, Philadelphia, Saunders, 1923.

modified series with a clear exponential trend may be secured. The trend of the original series may be written

$$y = K + ar^x$$

where K is the constant magnitude by which the series departs from a geometric progression. A *modified exponential* curve of this type may give a highly satisfactory representation of trend, in certain cases. The method employed in fitting such a curve is discussed in Appendix D.

Some use has been made, in the interpretation of economic statistics, of the Gompertz curve, the equation to which was originally developed in the actuarial field. The equation is

$$y = ab^{c^x}.$$

Its use in the analysis of economic statistics has been based upon the argument that there is a general law of growth characteristic of population increase, and that this same type of growth is found in industries whose products are a direct function of the growth of population.

A somewhat similar curve of growth, the "logistic," has been employed by Verhulst and more recently by Raymond Pearl and Lowell J. Reed in forecasting population growth. This curve has been found to describe the trends of certain economic series. Examples of the procedures employed in fitting Gompertz and logistic curves are given in Appendix D.

The Determination of Monthly Trend Values

The procedures so far described have dealt with annual measurements only. Having fitted a line or curve to annual data it is frequently necessary to make a transition to monthly units. Problems involving such monthly measurements are faced in the study of cyclical movements which are discussed in the next chapter.

The constant a in the trend equation defines the trend value in the year taken as origin. If the annual data

employed in the fitting processes are averages of twelve monthly values (e.g., the average price of pig iron in 1937) the constant a measures the trend value for a month centering at the middle of the year covered by the annual figures. If the annual data are aggregates of twelve monthly values (e.g., total production of pig iron in 1937) the constant a must be divided by 12 to obtain the trend value for the month centering at the middle of the year.

If the trend be linear, the constant b in the equation $y = a + bx$ defines the change due to trend over a twelve-month period. In interpolating for monthly trend values, the increment (or decrement) from month to month (e.g., from January to February of the year 1937) is $\frac{b}{12}$, if the annual data employed in the fitting process are averages of monthly values. The increment from month to month is $\frac{b}{144}$ if the annual data are aggregates of monthly values.

The one further step needed is properly to center the monthly trend values. These should, of course, be centered at points of time corresponding to those to which the actual monthly data relate. In averaging, or aggregating, monthly data relating to the middle of each of the twelve months in a calendar year we secure a figure centering at July 1. The month centering at the middle of the year of origin thus centers at July 1. For comparison with actual monthly data, we desire trend values centering at July 15, August 15, etc. At the beginning, therefore, we must add to the trend value for the month centering at the middle of the year of origin $\left(\text{that is, to } a \text{ or to } \frac{a}{12}\right)$ one-half of the month-to-month increment (or decrement) that we have obtained from b of the trend equation. This procedure gives us the trend value for the month centering at July 15. This value may be compared with the actual value recorded for that month. The addition to this of the month-to-month trend

increment (or decrement) gives trend values for all following months; subtraction gives trend values for all preceding months.[1]

The Selection of a Curve to Represent Trend

Various types of curves which may be fitted to represent the trend of economic data over a period of time have been described. But which of these many types is to be selected in a given case? Which will give the best standard of normality for each of the years covered? Several references to this problem have been made in the preceding sections, but no general principles have been laid down. And, in fact, no general principles can be evoked to answer this fundamental question. There is no absolute test of goodness of fit in such cases. It is largely a matter of personal judgment as to the type of curve which best represents the trend in a given instance, and experience must play a dominant part in such judgments. But certain general considerations are of assistance in selecting the appropriate type of curve.

1. The first step in the selection of a curve type is the plotting of the data. When this has been done, it is frequently possible by inspection to determine the appropriate form. The data may be plotted in four different combinations, of which the first two are of chief importance in dealing with economic material.

a. Natural x, natural y. (That is, plot the given figures on ordinary arithmetic paper.)

b. Natural x, log y. (Plot the x's on the natural scale, and plot the y's on the logarithmic scale; i.e., use semi-logarithmic paper.)

[1] If the original monthly data relate to the first or last of the month, rather than the middle, a similar correction is needed, but the monthly dates named in the text would be different, of course. If the trend equation is non-linear, the process of interpolation must be correspondingly modified. For a discussion of appropriate procedures the reader is referred to any treatise dealing with the general principles of interpolation. *The Calculus of Observations*, by Whitaker and Robinson, contains an excellent treatment of this topic.

c. Natural y, log x. (Plot on semi-logarithmic paper, with the x-scale logarithmic.)

d. Log y, log x. (Plot on paper with logarithmic ruling on both scales.)

If in any of these cases a straight line trend is secured, a type of equation which plots as a straight line under the given conditions (cf. Chapter II) would be selected. If a linear equation is not appropriate some other simple type may be suggested by the plotted data. In studying such graphs for the purpose of selecting a curve to represent trend, one should be familiar with the curves representing all the simpler equations.

2. The appropriate curve may be determined by a study of the relations between the two variables, x and y. In the simpler cases the following relations hold:[1]

a. If, when the values of x are arranged in an arithmetic series, the corresponding values of y form a geometric series, the relation is of the exponential type, described by the equation

$$y = ab^x.$$

b. If, when the values of x are arranged in a geometric series, the corresponding values of y form a geometric series, the relation is of the simple parabolic or hyperbolic type, described by the equation

$$y = ax^b.$$

c. If, when the values of x are arranged in an arithmetic series, the first differences of the corresponding y's are constant, the relation is of the straight line type, described by the equation

$$y = a + bx.$$

The differences between successive y values, when x's are arranged in an arithmetic series, are termed "first differences" or "first order differences" and are represented by the symbol Δy. The differences between successive first differences are called "second differences" and are represented by the

[1] It will be recalled that an arithmetic series changes by a constant absolute increment, while a geometric series changes by a constant percentage.

symbol $\Delta^2 y$. Differences of higher order are similarly derived. The following table illustrates the formation of differences:

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
1	11			
		29		
2	40		32	
		61		12
3	101		44	
		105		12
4	206		56	
		161		12
5	367		68	
		229		12
6	596		80	
		309		12
7	905		92	
		401		12
8	1,306		104	
		505		12
9	1,811		116	
		621		
10	2,432			

d. If, when the values of x are arranged in an arithmetic series, the nth differences of the corresponding y's are constant, the relation between the variables is described by an equation of the potential series carried to the nth power of x; that is, by an equation of the type

$$y = a + bx + cx^2 + dx^3 + \ldots + qx^n.$$

Thus, in the example given above, in which the third differences are constant, the relation between x and y would be described by an equation of the form

$$y = a + bx + cx^2 + dx^3.$$

When one is selecting a curve to use in the analysis of economic data, he will rarely, if ever, find these tests to be met perfectly. This would happen only when the curve chosen passed through all the plotted points. But data in a given case will generally approximate some one of the conditions described above, and the appropriate type of curve will be indicated.

3. If the study of the original data does not render a definite decision possible, several types of curves may be fitted to the data and the decision made by comparing the results. If the equations to the curves being compared contain the same number of constants, a comparison of the root-mean-square deviations about the curves furnishes a conclusive and valid test of the closeness of the fit within the limits of the data.

The root-mean-square deviation may be readily computed by making use of the following relationship

$$\Sigma(d^2) = \Sigma(y^2) - a\Sigma(y) - b\Sigma(xy) - c\Sigma(x^2y) - \ldots$$

where $\Sigma(d^2)$ is the sum of the squares of the deviations about the line of trend. (The derivation of this equation is explained in Appendix A, in which a generalized form is given.) If the equations do not contain an equal number of constants, a test of this sort is invalid and the comparison can only be made by inspection. Personal judgment as to the curve which represents the trend most accurately must be the basis of the decision in such cases.

It should be remembered that the closeness of fit within the limits of the data is not of itself a final criterion. An equation could be secured, having a number of constants equal to the number of points, which would give a curve passing through every point plotted, yet such a curve would not necessarily represent the trend. The concept of a *trend* is of a regular, smooth underlying movement, from which there are deviations, but which marks the long-time tendency of the series. In general, therefore, the curve should be of simple form, if it is to be consistent with the concept of secular trend. This does not mean, however, that a complex trend can be represented by a simple curve which fails to conform to the plotted data.

4. An important question to be answered before the form of curve can be selected relates to the limits within which the line of trend is to be used. If it is to be used only within the limits of the plotted data (i.e., for *interpolation*) one set of considerations governs the choice of a curve. If it is to be projected beyond the limits of the data, used as a basis for the determination of *normal* during a subsequent period, other considerations enter. In the former case a reasonable fit to the data is the sole requirement; in the latter case it is necessary, in addition, that the trend of the projection be logical, and consistent with the past record.

The fact should be clearly recognized that projection, or *extrapolation*, represents a guess, justified only on the assumption that a proper line of trend has been fitted and that the same conditions that affected the series in the past will prevail in the future. A change in conditions, the introduction of new elements, renders the projection invalid. When dealing with economic statistics, moreover, it is ordinarily impossible to tell, except in retrospect, when a change has taken place. Conclusions drawn from the projection of a line of trend are always subject to error, therefore. In practical statistical work such projections are made, and are justified on the ground that the most probable course in the future is that which prevailed in the past. Projections into the distant future are, of course, subject to wider margins of error than short-time projections. Lines of trend should be revised from time to time, therefore, as new data become available.

When a projection is to be made, a simple curve with few constants is to be preferred to a more complicated one. A third or fourth degree parabola may give an excellent fit to the data in a given case, but the projection of such curves is inadvisable. It is well to remember, as Perrin has pointed out, that a curve suitable for interpolation may not be at all adapted to extrapolation.

The avoidance of distortion of trend lines by abnormal conditions in the terminal years of the period studied is particularly important when a trend is to be projected. Reference is made to this point in the next chapter.

It seems to be true, in general, that simple curves fitted to the logarithms of the y's give more reliable results when projected than curves fitted to the natural numbers. In an interesting discussion of this point, Karl G. Karsten[1] argues that phenomena characterized by a uniform *rate of change* are more likely to maintain their trend than phenomena marked by a uniform *amount of change*. It is

[1] Karl Karsten, *Charts and Graphs*, New York, Prentice Hall, 1923, 423–425.

the semi-logarithmic curves, of course, which best measure *rates* of change.

5. It is frequently true that no one curve will fit a given series during the entire period it is desired to study. This may be due to changes in conditions which cause the trend to be altered. Thus the trend of wholesale prices was downward, in a direction well represented by a straight line, from the close of the Civil War to 1896. From 1896 to the beginning of the World War the trend was upward, and could be described by a second degree parabola. From 1921 to 1929 the trend was also curvilinear, rising to 1925, declining thereafter. Similar changes occur in many economic series. By breaking the entire period up into sections, appropriate lines of trend may be fitted to the several periods thus marked off. This process may be carried to a quite illogical extreme, however. The concept of trend is of a gradual, long-term change, and the breaking up of a series in order to fit a number of trend lines is contrary to the whole conception. It may be justified upon occasion when a real change in conditions occurs, but in all cases the attempt should be made to represent the trend during the whole period by a single line.

Deflation as a Step in Analysis

Many series of economic data are expressed in monetary units, in dollars, pounds, or francs. Such series are subject to distortion because of changes in the price level. Thus the value of heavy engineering contracts awarded in the United States in 1913 amounted to approximately 601 millions of dollars; in 1929 the value of engineering contracts awarded in the same territory amounted to approximately 3,950 millions of dollars.[1] Was the volume of engineering construction in 1929 over six times that in 1913? It was not. The value of construction contracts awarded in any year depends not only upon the actual volume of construc-

[1] Figures compiled by *Engineering News Record.*

tion but also upon the costs of construction materials and labor, and these costs increased substantially from 1913 to 1929. If we wish to measure the change in the volume of construction alone, these values must be corrected for the increase in construction costs between 1913 and 1929. Such a process is termed *deflation.*[1]

The selection of an appropriate deflating index is the central problem in such cases. For the present purpose we

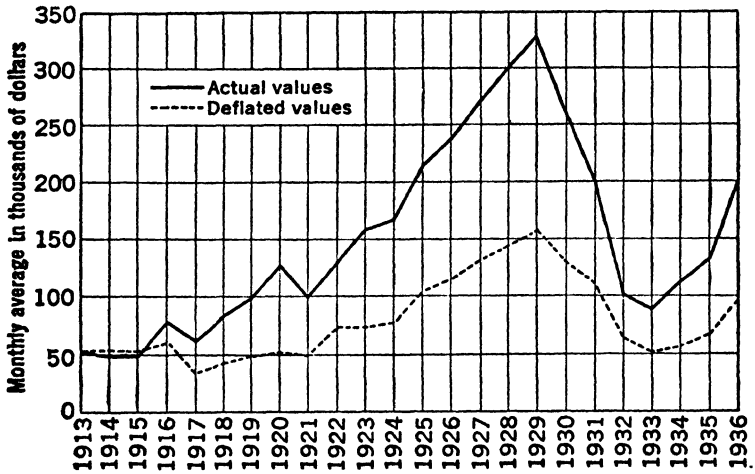


FIG. 61. — Comparison of Actual and Deflated Values of Contracts Awarded in Engineering Construction, 1913–1936

may use an index of constructive costs, based upon the prices of steel, cement, and lumber, and upon wage rates in construction industries, compiled by the *Engineering News Record.* This index shows that construction costs in 1929 were approximately 107 per cent higher than in

[1] The term *deflation* is not inappropriate when correction is being made for an advance in prices; it is less suitable when correction is made for a fall in prices. The period selected as a standard of reference may be one in which prices were relatively high; division by a price or cost index resting on such a year as base will *raise* values relating to other periods. The word *deflation* is a convenient one to use for this general process, however. In using it in this somewhat technical sense we must understand it to mean *correction for changes in the value of the dollar* (as measured by specific indices of prices or costs).

TABLE 74

Actual and Deflated Values of Contracts Awarded in Engineering Construction, 1913–1936

Year	*Contracts awarded, engineering construction (monthly average, in thousands of dollars)* [1]	*Index of construction costs* [1]	*Deflated value of contracts awarded (monthly average, in thousands of dollars)*
1913	50,117	1.000	50,117
1914	48,574	.886	54,824
1915	48,740	.926	52,635
1916	77,778	1.296	60,014
1917	61,592	1.812	33,991
1918	82,729	1.892	43,726
1919	97,991	1.984	49,391
1920	126,923	2.513	50,507
1921	99,459	2.018	49,286
1922	129,716	1.745	74,336
1923	158,670	2.141	74,110
1924	166,593	2.154	77,341
1925	213,287	2.067	103,187
1926	237,820	2.080	114,337
1927	271,147	2.062	131,497
1928	298,215	2.068	144,205
1929	329,193	2.070	159,030
1930	264,438	2.029	130,329
1931	202,693	1.814	111,738
1932	101,609	1.570	64,719
1933	89,031	1.702	52,310
1934	113,383	1.981	57,235
1935	132,513	1.952	67,886
1936	198,904	2.065	96,322

1913 (the index is 100 for 1913, 207 for 1929). Dividing the 1929 aggregate by 2.07, to correct for the change in costs, we secure a *deflated* total of 1,908 millions of dollars. This may be taken to measure the aggregate value of engineering contracts awarded in 1929, when the 1913 dollar is used as a standard of value. (In this process the value of money is assumed to be held constant with ref-

[1] Data on contracts awarded have been compiled by the *Engineering News Record;* the index of construction costs has been computed by the same agency.

erence to the year which is the base of the price or cost index used as a deflator.) If the deflating index may be accepted as an accurate measure of changing costs, the deflated series may be assumed to define changes in the actual volume of engineering construction. The effects of changing prices and wages will have been eliminated.

The general procedure is illustrated in greater detail in Table 74 on page 281. Actual and deflated series are plotted in Fig. 61. The degree to which changing monetary values distorted the construction series may be readily appreciated from the diagram.

Most value series are affected by price changes, and it is generally advisable to correct for this factor before further analysis is attempted. Each case presents a new problem, for no general deflating index is suitable to all series. The index of wholesale prices compiled by the United States Bureau of Labor Statistics has been used extensively in deflating economic data expressed in dollar values, but this index is not at all appropriate in many of the cases in which it has been employed. It is absurd, for instance, to deflate money wages by an index of wholesale prices. The deflating index employed should be a measure of price changes as they affect the series being deflated.

The deflation of a value series is in general a first step in the study of that series. The way is then open for further analysis by methods explained in the present and succeeding chapters.

REFERENCES

Bowley, A. L., *Elements of Statistics*, Chap. 7.

Burns, Arthur F., *Production Trends in the United States Since 1870*, Chap. 2.

Camp, B. H., *The Mathematical Part of Elementary Statistics*, Part I, Chap. 7.

Chaddock, R. E., *Principles and Methods of Statistics*, Chap. 13.

Croxton, F. E. and Cowden, D. J., *Practical Business Statistics*, Chap. 15.

Crum, W. L. and Patton, A. C., *An Introduction to the Methods of Statistics*, Chap. 20.

Davies, G. R. and Crowder, W. F., *Methods of Statistical Analysis in the Social Sciences*, Chap. 6.

Davies, G. R. and Yoder, Dale, *Business Statistics*, Chap. 4.

Day, E. E., *Statistical Analysis*, Chaps. 15–17.

Frickey, Edwin, "The Problem of Secular Trend," *Review of Economic Statistics*, Oct. 15, 1934.

Kuznets, Simon, *Secular Movements in Production and Prices.*

Kuznets, Simon, "Time Series," *Encyclopædia of the Social Sciences*, Vol. 14.

Lipka, Joseph, *Graphical and Mechanical Computation.*

Macaulay, F. R., *The Smoothing of Time Series.*

Mills, F. C., *Economic Tendencies in the United States.*

Mitchell, W. C., *Business Cycles, The Problem and Its Setting*, Chap. 3.

Pearl, Raymond, *Medical Biometry and Statistics*, Chap. 16.

Rhodes, E. C., *Elementary Statistical Methods*, Chap. 12.

Richardson, C. H., *An Introduction to Statistical Analysis*, Chaps. 6, 8.

Running, T. R., *Empirical Formulas.*

Sasuly, Max, *Trend Analysis of Statistics.*

Smith, J. G., *Elementary Statistics*, Part III, Chaps. 11, 12.

Steinmetz, C. P., *Engineering Mathematics*, Chap. 6.

Waugh, A. E., *Elements of Statistical Method*, Chap. 8.

CHAPTER VIII

THE ANALYSIS OF TIME SERIES: MEASUREMENT OF SEASONAL AND CYCLICAL FLUCTUATIONS

The measurement of secular trend is but one of the problems connected with the analysis of a series in time. Such series, it has been pointed out, are subject to periodic fluctuations, seasonal and cyclical in character, and these fluctuations are generally more important in their effects upon business than is the long-time trend. Our present concern is with methods of isolating such periodic variations. The series, in Table 75, which clearly reflects the seasonal and cyclical swings of domestic trade in the United States, may be used to illustrate methods of measuring these movements.

TABLE 75

Average Weekly Freight Car Loadings in the United States, 1918–1927 [1]

(Unit: 1,000 cars)

Month	1918	1919	1920	1921	1922	1923	1924	1925	1926	1927
January	655	728	820	706	696	848	859	891	920	944
February	753	687	776	685	757	854	908	906	932	956
March	842	696	848	691	818	916	916	926	960	998
April	873	721	730	706	716	941	874	932	966	969
May	897	759	862	760	776	975	895	971	1,018	1,004
June	918	796	896	762	831	1,012	906	992	1,052	1,021
July	970	887	901	750	813	985	881	975	1,037	978
August	962	892	969	810	853	1,042	969	1,073	1,106	1,073
September	956	960	967	842	925	1,037	1,037	1,074	1,140	1,093
October	925	967	1,005	932	978	1,070	1,091	1,107	1,184	1,101
November	819	807	884	764	957	964	976	1,024	1,042	926
December	719	758	755	681	832	827	869	925	858	814
Average	857	805	868	757	829	956	932	983	1,018	990

[1] Data from the *Annual Bulletin* of the American Railway Association and the *Survey of Current Business*. The published figures have been slightly revised, to take account of calendar variations.

For the present purpose the study of seasonal and cyclical variations in freight car loadings is limited to the period 1918–1927. The disturbances of the ensuing period, combined with changes in railroad operating methods and business practices, materially modified the behavior of this series. The demonstration of statistical procedure will be clearer if restricted to the relatively homogeneous period here covered.

The Measurement of Seasonal Fluctuations: Moving Averages

Moving averages provide a useful method of defining seasonal variations. Since these fluctuations take place within a constant period of twelve months, a moving average may be used with more confidence than when a cycle of varying length is involved. The magnitude of the fluctuations (the *amplitude* of the seasonal swings) will not ordinarily be constant, hence the line marked out by the moving averages will not be completely free of seasonal influences. The relation of the actual monthly items to the moving averages may be averaged, however, and the indices of seasonal variation based upon these averages.

It is essential, of course, that the moving average, centered, fall at the same date as the original figure with which it is to be compared. This involves a second process of averaging. For example, the weekly averages of freight car loadings relate to the middle of each month. The average of the twelve monthly items for 1918, when centered, falls on July 1st. The average of the items from February, 1918, through January, 1919, centered, falls on August 1st. To secure a figure comparable with the July 15th average, these two must be averaged. By this process of computing a two-month moving average from the twelve-month average, comparability with the original figures is secured. Table 76 presents averages obtained in this way for the period from July, 1918, to June, 1927.

TABLE 76

Moving Averages of Freight Car Loadings, 1918–1927

(12-month moving average, centered, adjusted by 2-month moving average centered)

(Unit: 1,000 cars)

Month	1918	1919	1920	1921	1922	1923	1924	1925	1926	1927
Jan.		808.0	850.8	809.6	783.7	915.8	935.9	957.3	1,004.8	1,019.1
Feb.		801.7	854.6	796.7	788.1	930.9	928.5	965.6	1,008.7	1,015.3
Mar.		798.9	858.1	784.9	793.4	943.4	925.5	971.5	1,012.8	1,012.0
Apr.		800.8	860.0	776.8	798.8	951.9	926.4	973.7	1,018.8	1,006.5
May		802.1	864.8	768.8	808.7	956.0	927.8	976.3	1,022.8	998.3
June		803.2	867.9	760.5	823.0	956.1	930.0	980.7	1,020.7	991.6
July	860.5	808.7	863.0	757.0	835.7	956.4	933.1	984.2	1,018.9	
Aug.	860.8	816.2	854.5	759.6	846.0	959.1	934.3	986.5	1,020.9	
Sept.	851.9	826.3	844.1	767.9	854.2	961.3	934.7	989.0	1,023.5	
Oct.	839.5	833.0	836.6	773.6	867.6	958.5	937.5	991.8	1,025.2	
Nov.	827.4	837.6	831.3	774.7	885.3	952.4	943.1	995.2	1,024.8	
Dec.	816.6	846.1	821.5	778.2	901.1	944.7	949.8	999.7	1,022.9	

The original data are now expressed as percentages of the corresponding moving averages. These percentages are given in Table 77.

TABLE 77

Percentage Relation of Actual Freight Car Loadings to 12-Month Moving Averages

Month	1918	1919	1920	1921	1922	1923	1924	1925	1926	1927
Jan.		90.1	96.4	87.2	88.8	92.6	91.8	93.1	91.6	92.6
Feb.		85.7	90.8	86.0	96.1	91.7	97.8	93.8	92.4	94.2
Mar.		87.1	98.8	88.0	103.1	97.1	99.0	95.3	94.8	98.6
Apr.		90.0	84.9	90.9	89.6	98.9	94.3	95.7	94.8	96.3
May		94.6	99.7	98.9	96.0	102.0	96.5	99.5	99.5	100.6
June		99.1	103.2	100.2	101.0	105.8	97.4	101.2	103.1	103.0
July	112.7	109.7	104.4	99.1	97.3	103.0	94.4	99.1	101.8	
Aug.	111.8	109.3	113.4	106.6	100.8	108.6	103.7	108.8	108.3	
Sept.	112.2	116.2	114.6	109.6	108.3	107.9	110.9	108.6	111.4	
Oct.	110.2	116.1	120.1	120.5	112.7	111.6	116.4	111.6	115.5	
Nov.	99.0	96.3	106.3	98.6	108.1	101.2	103.5	102.9	101.7	
Dec.	88.0	89.6	91.9	87.5	92.3	87.5	91.5	92.5	83.9	

These percentages show some variation from year to year in the relation of the figures for a given month to the moving average. Thus the January figures, while always below the average, vary from 87.2 per cent to 96.4 per cent of the average. The nine percentages secured for each month must be averaged to obtain the index

desired. Either the arithmetic average or the median may be employed for this purpose. The results secured by applying the two methods are shown in Table 78. In columns (2) and (3) the actual arithmetic means and medians are given. The average of the twelve arithmetic means happens to be exactly 100, so no further adjustment is needed. Usually the average will depart in some degree from 100, as it does for the medians. When this is the case, the twelve monthly index numbers must be adjusted to make their average equal to 100. The items in column (4) have been secured from the items in column (3) by dividing throughout by 1.00367.

TABLE 78

Indices of Seasonal Variation in Freight Car Loadings, Computed from Moving Averages

(1) *Month*	(2) *Arithmetic means*	(3) *Medians (unadjusted)*	(4) *Medians (adjusted)*
January	91.6	91.8	91.5
February	92.1	92.4	92.1
March	95.8	97.1	96.7
April	92.8	94.3	94 0
May	98.6	99.5	99.1
June	101.6	101.2	100.8
July	102.4	101.8	101.4
August	107.9	108.6	108.2
September	111.1	110.9	110.5
October	115.0	115.5	115.1
November	101.7	101.7	101.3
December	89.4	89 6	89.3
Average	100.0	100.367	100.0

THE COMPUTATION OF INDEX NUMBERS OF SEASONAL VARIATION BY AVERAGING RATIOS TO TREND

A somewhat similar method of securing seasonal indices, which has certain distinctive advantages, involves the averaging of ratios to trend.[1] In the application of this

[1] The essentials of this method were worked out independently by Helen D.

method, a suitable line of trend, linear or non-linear, is fitted to the data, the actual monthly items are expressed as percentages of the corresponding trend figures, and then, for each month, an average of the percentage ratios of the actual to the trend values is secured. This procedure is identical with that described in connection with the use of moving averages, except that the actual values may be expressed as percentages of normal values derived from any function employed to represent trend. In the selection of an average value for each month, use may be made of a multiple frequency table in obtaining an understanding of the nature of the actual seasonal movement. With the help of such a table the existence of a definite seasonal movement may be verified and the type of average to be used in securing a typical value for each month may be determined. (It would, of course, be equally appropriate to use a table of this type in connection with the method of moving averages.) We shall apply this method to the data employed in the preceding examples.

A straight line, fitted to annual averages of the data of freight car loadings from 1918 to 1927, as given in Table 75, is described by the equation

$$y = 769.00 + 23.727x$$

with origin at July 1, 1917. Normal values for each month may be computed readily.[1] The normal value for the month centering at July 1, 1917, is 769.00 (i.e., the constant a of the trend equation). Since the increment over a twelve-month period is 23.727, the increment from month to month is one twelfth of this, or, 1.977. Hence the normal value for the month centering at January 1, 1918, is 769.00

(*Footnote 1 continued from page 287.*)
Falkner "The Measurement of Seasonal Variation," *Journal of the American Statistical Association*, June, 1924, 167–179, and Lincoln W. Hall, "Seasonal Variation as a Relative of Secular Trend," *Journal of the American Statistical Association*, June, 1924, 156–166.

[1] Methods used in the determination of monthly trend values are discussed in Chapter VII.

+ (6 × 1.977), or 780.862. But the average weekly freight car loadings for January, 1918, must be taken to center at January 15th. The monthly trend value centering at that date is 780.862 + ½(1.977), or 781.850. The trend value for February, 1918, is secured by adding to 781.850

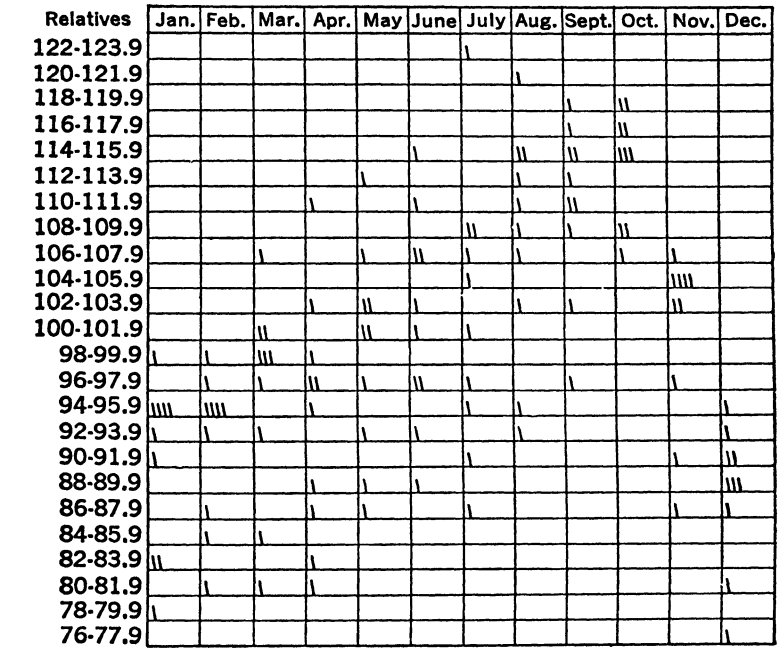


FIG. 62. — Frequency Distributions: Monthly Freight Car Loadings Expressed as Relatives of Corresponding Trend Values

the monthly increment, 1.977. A similar process gives the value for each succeeding month. The results, rounded off to the nearest whole number, are given in column (2) of Table 80.

Expressing each of the given values for each month as a percentage of the corresponding trend value, we secure ten such relative figures (since the data cover ten years). The ten January percentages vary from 79.4 to 98.9, the ten October percentages from 107.0 to 119.7, etc. The multiple frequency table which appears in Fig. 62 is constructed

by classifying, in the form of a frequency distribution, the items for each month.

The presence of a distinct seasonal variation is demonstrated by this table. Freight traffic is consistently low in the winter months. Activity is somewhat greater in the spring, and reaches a peak as a result of harvesting and other demands in the late summer and fall.

The tabular summary facilitates the selection of a type of average for the measurement of the seasonal movements. The median is likely to be unrepresentative; it is subject to material change in value by the addition or withdrawal of one or two entries, unless there is a definite concentration in the monthly frequency distributions. The arithmetic mean of all the items, on the other hand, may be unduly affected by exceptional cases. An alternative method is provided by the possibility of taking the arithmetic mean of the central items for each month. If an inspection of the multiple frequency table does not lead to an immediate decision as to which is the best type of average to employ in a given case, several index numbers may be worked out for each month, and a decision reached after a comparison of the results. (Indeed, since the determination of a typical value is a separate problem for each month, the method of averaging employed might vary from month to month.) In the present instance the seasonal variation is fairly regular, year after year. No great differences would appear in the results secured by averaging varying numbers of items. Index numbers based upon averages of the four central items for each of the twelve months are appropriate in this case. (In general, an average of three, four, or five central values is more likely to be stable and representative than either the median or an average of all the items for each month. The greater the concentration in the monthly frequency tables, the smaller the number of items upon which the index numbers may be based.)

The simple averages of the four central items constitute

the unadjusted index numbers given in Table 79. Correcting these so that the average of each group is equal to 100, we secure the adjusted index numbers presented in the same table. (These averages have been derived, not from the frequency distributions shown in Fig. 62, but from individual percentages defining the relation of actual to trend values.)

TABLE 79

Indices of Seasonal Variation in Freight Car Loadings, Based upon Percentage Ratios of Actual Values to Linear Trend Values

Month	*Unadjusted index numbers (based upon four central items)*	*Adjusted index numbers (based upon four central items)*
January	92.9	91.6
February	94.8	93.5
March	98.6	97.2
April	94.3	93.0
May	100.2	98.8
June	102.4	101.0
July	102.8	101.3
August	109.7	108.2
September	112.3	110.7
October	115.6	114.0
November	103.6	102.1
December	89.9	88.6
Average	101.425	100.0

The index numbers of seasonal variation derived from ratios to trend accord very closely with those computed from moving averages. The widest discrepancy, for the month of February, amounts to only 1.4. The consistency of the seasonal movement in freight car loadings helps to explain this close agreement. In general the two methods here exemplified will yield results that are fairly close together. Both are well adapted to the measurement of seasonal changes in homogeneous series. Simpler methods may be used on occasion, and more involved methods may

be required in dealing with non-homogeneous series where there is reason to suspect that the pattern of seasonal movements has been modified during the period under observation.

Modifications of these general procedures are necessary when the pattern of the seasonal movements in a given series is altered during the period under observation. Two types of shifts in seasonal variation may be distinguished. The first includes shifts that are irregular over time, but that are related to definable causal factors. Thus the price of an agricultural product may follow one seasonal pattern in years of high production, and quite a different pattern in years of low production.[1] Where this condition prevails it may be possible to compute two sets of seasonal indices, each to be applied under appropriate conditions. Methods already described may be used in the construction of such indices. Of this irregular type, also, are alterations in the seasonal pattern of an economic series that reflect sharp changes in business practice. Shifts in the dates of the annual automobile shows in the United States have materially altered the seasonal index of automobile sales.

The second type of seasonal modification is progressive in character. The change in pattern is not sudden, nor does it reverse itself. Slow alterations over time in trade practices and consumption habits bring such evolutionary or secular changes. The slow displacement of the open car by the closed car brought such a progressive modification in the seasonal pattern of automobile sales. In the computation of seasonal indices under these conditions persistent changes over time in the figures for each month may be measured separately. Thus, when ratios to trend have been obtained, all the January items (such as those plotted in Fig. 62) may be plotted chronologically. The progressive change in the January relatives from 1920 to 1937, say, is then defined by a line of secular trend. The trend value

[1] See F. L. Thompson, *Agricultural Prices*, New York, McGraw-Hill, 1936.

for January of 1920 is a first approximation to the January seasonal index for 1920. The figure for February of 1920 is obtained in the same way, and so for each month of 1920. Adjustment of these preliminary values to make their average equal to 100 gives a set of seasonal indices for 1920. Seasonal indices for other years are computed in the same way.

This method is, of course, more laborious than the procedure followed when the seasonal pattern remains constant. Before applying the more complicated method the investigator should assure himself that the shift in pattern is real, and not merely a reflection of accidental variations.[1]

The Measurement of Cyclical Fluctuations

There remains the task of combining the corrections for secular trend and seasonal variation in order to secure measures of cyclical changes in a given series. Major interest in most economic studies attaches to these cyclical changes, and the measurement of such changes is usually the central problem in the analysis of time series. The complete elimination of all non-cyclical movements is impossible, of course. We must content ourselves with measures reflecting cyclical changes intermingled in rather uncertain proportions with accidental fluctuations.

The procedure may be illustrated with reference to the data of freight car loadings in the United States, presented in Table 75. For the purposes of the present illustration the study will be restricted to the decade 1918–1927. The

[1] Tests of sampling errors are discussed in Chapters XIV, XV, and XVIII. The test of a linear trend in this case would relate to the slope b of the line fitted to the relatives for a given month.

The literature on the measurement of seasonal fluctuations is extensive. The references at the close of this chapter contain detailed accounts of various modifications of the basic procedures discussed above. A rapid, flexible and accurate graphic method, suitable for use by the student who has grasped the essentials of the formal procedures, is explained in the article by William A. Spurr. Spurr's method utilizes relative (logarithmic) deviations, a procedure for which there is strong logical justification.

severe disturbances that occurred during the business cycle that ran its course between 1927 and 1933, and in the years immediately following, greatly complicate the task of disentangling the secular, seasonal, and cyclical elements in the behavior of this series. Not until a somewhat longer period has intervened will it be possible to determine the contributions a changing secular trend and changing seasonal movements may have made to the fluctuations in railway freight traffic during the decade 1927–1937.

In attempting to separate the results of secular, seasonal, cyclical, and random movements in the behavior of time series, it is well to establish a series of "expected" values, representing results of the operation of regularly acting forces. Most regular and predictable of the forces affecting time series are those defined as secular and seasonal. The equation to the line of secular trend of freight car loadings provides a means of estimating annual and monthly values. These would be the "expected" values were the forces of trend alone in operation. But we know that a seasonal movement, regular enough for fairly exact measurement, is superimposed upon the trend. The combination of the results of these two forces provides a basic series of "expected values," from which deviations due to the play of other forces may conveniently be measured.

A process suitable to this purpose is illustrated in Table 80. In col. (2) we have the monthly trend values of freight car loadings, and in col. (3) index numbers of seasonal variation. The products of the two, constituting the series of "expected values," are given in col. (4). Thus, for January, 1918, the expected number of freight cars loaded is not 782, the trend value, but $782 \times .916$, the latter figure being the seasonal index for January. This correction gives an "expected" number of 716. Subtracting from the actual values in col. (5) the corresponding expected values, we obtain the measurements in col. (6). The 655 cars loaded in January, 1918, fell short by 61 of the "expected"

TABLE 80

Illustrating the Analysis of a Series in Time

Freight Car Loadings, 1918–1927

(1) *Year and month*	(2) *Trend value* T	(3) *Seasonal index (as ratio)* S	(4) *Trend corrected for seasonal* TS	(5) *Actual value* A	(6) *Deviation of actual value from 'trend corrected for seasonal'* $A - TS$	(7) *Percentage deviation of actual value from 'trend corrected for seasonal'* $\frac{A - TS}{TS}$
1918						
Jan.	782	.916	716	655	− 61	− 8.5
Feb.	784	.935	733	753	+ 20	+ 2.7
Mar.	786	.972	764	842	+ 78	+ 10.2
Apr.	788	.930	733	873	+ 140	+ 19.1
May	790	.988	781	897	+ 116	+ 14.9
June	792	1.010	800	918	+ 118	+ 14.8
July	794	1.013	804	970	+ 166	+ 20.6
Aug.	796	1.082	861	962	+ 101	+ 11.7
Sept.	798	1.107	883	956	+ 73	+ 8.3
Oct.	800	1.140	912	925	+ 13	+ 1.4
Nov.	802	1.021	819	819	0	0
Dec.	804	.886	712	719	+ 7	+ 0.98
1919						
Jan.	806	.916	738	728	− 10	− 1.4
Feb.	808	.935	755	687	− 68	− 9.0
Mar.	810	.972	787	696	− 91	− 11.6
Apr.	812	.930	755	721	− 34	− 4.5
May	813	.988	803	759	− 44	− 5.5
June	815	1.010	823	796	− 27	− 3.3
July	817	1.013	828	887	+ 59	+ 7.1
Aug.	819	1.082	886	892	+ 6	+ 0.7
Sept.	821	1.107	909	960	+ 51	+ 5.6
Oct.	823	1.140	938	967	+ 29	+ 3.1
Nov.	825	1.021	842	807	− 35	− 4.2
Dec.	827	.886	733	758	+ 25	+ 3.4
1920						
Jan.	829	.916	759	820	+ 61	+ 8.0
Feb.	831	.935	777	776	− 1	− 0.1
Mar.	833	.972	810	848	+ 38	+ 4.7
Apr.	835	.930	777	730	− 47	− 6.0
May	837	.988	827	862	+ 35	+ 4.2
June	839	1.010	847	896	+ 49	+ 5.8

TABLE 80—*Continued*

Illustrating the Analysis of a Series in Time

(1)	(2)	(3)	(4)	(5)	(6)	(7)
	T	S	TS	A	$A - TS$	$\frac{A - TS}{TS}$
1920						
July	841	1.013	852	901	+ 49	+ 5.8
Aug.	843	1.082	912	969	+ 57	+ 6.3
Sept.	845	1.107	935	967	+ 32	+ 3.4
Oct.	847	1.140	966	1,005	+ 39	+ 4.0
Nov.	849	1.021	867	884	+ 17	+ 2.0
Dec.	851	.886	754	755	+ 1	+ 0.1
1921						
Jan.	853	.916	781	706	− 75	− 9.6
Feb.	855	.935	799	685	− 114	− 14.3
Mar.	857	.972	833	691	− 142	− 17.0
Apr.	859	.930	799	706	− 93	− 11.6
May	861	.988	851	760	− 91	− 10.7
June	863	1.010	872	762	− 110	− 12.6
July	865	1.013	876	750	− 126	− 14 4
Aug.	867	1.082	938	810	− 128	− 13.6
Sept.	869	1.107	962	842	− 120	− 12.5
Oct.	871	1.140	993	932	− 61	− 6.1
Nov.	873	1.021	891	764	− 127	− 14.3
Dec.	875	.886	775	681	− 94	− 12.1
1922						
Jan.	877	.916	803	696	− 107	− 13.3
Feb.	879	.935	822	757	− 65	− 7.9
Mar.	881	.972	856	818	− 38	− 4.4
Apr.	883	.930	821	716	− 105	− 12.8
May	885	.988	874	776	− 98	− 11.2
June	887	1.010	896	831	− 65	− 7.3
July	889	1.013	901	813	− 88	− 9.8
Aug.	891	1 082	964	853	− 111	− 11.5
Sept.	893	1.107	989	925	− 64	− 6.5
Oct.	895	1.140	1,020	978	− 42	− 4.1
Nov.	897	1.021	916	957	+ 41	+ 4.5
Dec.	899	.886	797	832	+ 35	+ 4.4
1923						
Jan.	900	.916	824	848	+ 24	+ 2.9
Feb.	902	.935	843	854	+ 11	+ 1.3
Mar.	904	.972	879	916	+ 37	+ 4.2
Apr.	906	.930	843	941	+ 98	+ 11.6
May	908	.988	897	975	+ 78	+ 8.7

TABLE 80—*Continued*

Illustrating the Analysis of a Series in Time

(1)	(2) T	(3) S	(4) TS	(5) A	(6) $A - TS$	(7) $\frac{A - TS}{TS}$
1923						
June	910	1.010	919	1,012	+ 93	+ 10.1
July	912	1.013	924	985	+ 61	+ 6.6
Aug.	914	1.082	989	1,042	+ 53	+ 5.4
Sept.	916	1.107	1,014	1,037	+ 23	+ 2.3
Oct.	918	1.140	1,047	1,070	+ 23	+ 2.2
Nov.	920	1.021	939	964	+ 25	+ 2 7
Dec.	922	.886	817	827	+ 10	+ 1.2
1924						
Jan.	924	.916	846	859	+ 13	+ 1.5
Feb.	926	.935	866	908	+ 42	+ 4.8
Mar.	928	.972	902	916	+ 14	+ 1.6
Apr.	930	.930	865	874	+ 9	+ 1.0
May	932	.988	921	895	− 26	− 2.8
June	934	1.010	943	906	− 37	− 3.9
July	936	1.013	948	881	− 67	− 7.1
Aug.	938	1.082	1,015	969	− 46	− 4.5
Sept.	940	1.107	1,041	1,037	− 4	− 0.4
Oct.	942	1.140	1,074	1,091	+ 17	+ 1.6
Nov.	944	1.021	964	976	+ 12	+ 1.2
Dec.	946	.886	838	869	+ 31	+ 3.7
1925						
Jan.	948	.916	868	891	+ 23	+ 2.6
Feb.	950	.935	888	906	+ 18	+ 2.0
Mar.	952	.972	925	926	+ 1	+ 0.1
Apr.	954	.930	887	932	+ 45	+ 5.1
May	956	.988	945	971	+ 26	+ 2.8
June	958	1.010	968	992	+ 24	+ 2.5
July	960	1.013	972	975	+ 3	+ 0.3
Aug.	962	1.082	1,041	1,073	+ 32	+ 3.1
Sept.	964	1.107	1,067	1,074	+ 7	+ 0.7
Oct.	966	1.140	1,101	1,107	+ 6	+ 0.5
Nov.	968	1.021	988	1,024	+ 36	+ 3.6
Dec.	970	.886	859	925	+ 66	+ 7.7
1926						
Jan.	972	.916	890	920	+ 30	+ 3.4
Feb.	974	.935	911	932	+ 21	+ 2.3
Mar.	976	.972	949	960	+ 11	+ 1.2
Apr.	978	.930	910	966	+ 56	+ 6.2
May	980	.988	968	1,018	+ 50	+ 5.2

TABLE 80—*Continued*

Illustrating the Analysis of a Series in Time

(1)	(2) T	(3) S	(4) TS	(5) A	(6) $A - TS$	(7) $\frac{A - TS}{TS}$
1926						
June	982	1.010	992	1,052	+ 60	+ 6.0
July	984	1.013	997	1,037	+ 40	+ 4.0
Aug.	986	1.082	1,067	1,106	+ 39	+ 3.7
Sept.	987	1.107	1,093	1,140	+ 47	+ 4.3
Oct.	989	1.140	1,127	1,184	+ 57	+ 5.1
Nov.	991	1.021	1,012	1,042	+ 30	+ 3.0
Dec.	993	.886	880	858	− 22	− 2.5
1927						
Jan.	995	.916	911	944	+ 33	+ 3.6
Feb.	997	.935	932	956	+ 24	+ 2.6
Mar.	999	.972	971	998	+ 27	+ 2.8
Apr.	1,001	.930	931	969	+ 38	+ 4.1
May	1,003	.988	991	1,004	+ 13	+ 1.3
June	1,005	1.010	1,015	1,021	+ 6	+ 0.6
July	1,007	1.013	1,020	978	− 42	− 4.1
Aug.	1,009	1.082	1,092	1,073	− 19	− 1.7
Sept.	1,011	1.107	1,119	1,093	− 26	− 2.3
Oct.	1,013	1.140	1,155	1,101	− 54	− 4.7
Nov.	1,015	1.021	1,036	926	− 110	− 10.6
Dec.	1,017	.886	901	814	− 87	− 9.7

number, 716. Such deviations of actual values from "trend corrected for seasonal" represent the combined influence of cyclical and accidental factors. These may be utilized in the absolute form given in col. (6), or may be expressed in percentage terms as in col. (7) of Table 80.

The series defining trend values corrected for seasonal variations, which are given in cols. (6) and (7) of Table 80, furnish the most satisfactory bases from which cycles in economic series may be measured. It is true that the "cycles" in cols. (6) and (7) are distorted by accidental fluctuations, but there is no simple means by which these may be eliminated. Recognizing their presence, the series may be put to fruitful use in the study of cyclical movements.[1]

[1] A series of "corrected deviations from trend" may be secured by subtracting the indices of seasonal variation from a series in which actual values are

The analysis of this series may be followed through graphically in Figs. 63 and 64 on page 300. The actual data of freight car loadings, by months from 1918 to 1927, are plotted in Fig. 63, together with a straight line of trend. In addition, a series of expected values (the figures in col. [4] of Table 80) is given for comparison with the actual. In this chart the seasonal pattern, shown by the dotted line, is superimposed upon the trend. Fig. 64 shows the deviations of actual from expected values, in percentage terms. These constitute the "cycles" in freight car loadings. As we have noted, random elements as well as cyclical fluctuations proper are present in these deviations. It would be possible, by using three- or five-month moving averages on these deviations, or by other smoothing processes, to eliminate some of the effects of the accidental movements. But the random and the cyclical movements are so closely interwoven that the attempt at separation is not generally made.

If cyclical changes in this series are to be compared with similar changes in other series, it is desirable to reduce the figures to a form permitting such comparison. The percentage deviations might be much more violent in one series than in another, and without a common denominator comparison would be difficult. This common denominator is afforded by the standard deviation. The monthly or annual deviations may be expressed in terms of the standard deviation as the unit of measurement, if such comparison is to be made.

(*Footnote 1 continued from page 298.*)

given as percentages of corresponding trend values. That is, $\frac{A}{T} - S$ may be employed, instead of $\frac{A - TS}{TS}$. This usage, which involves the assumptions that the "cyclical-accidental" composite and seasonal variations both represent deviations from trend as base and that their influences are additive, is not as strong, logically, as the method exemplified in the text. Trend and seasonal forces are the constant factors in the behavior of time series. In combination they may be thought of as providing the base from which cyclical and accidental movements occur, as deviations. (This is a convenient, and perhaps not a faulty, conception. We do not, however, possess knowledge of the true organic relations among the elements of time series.)

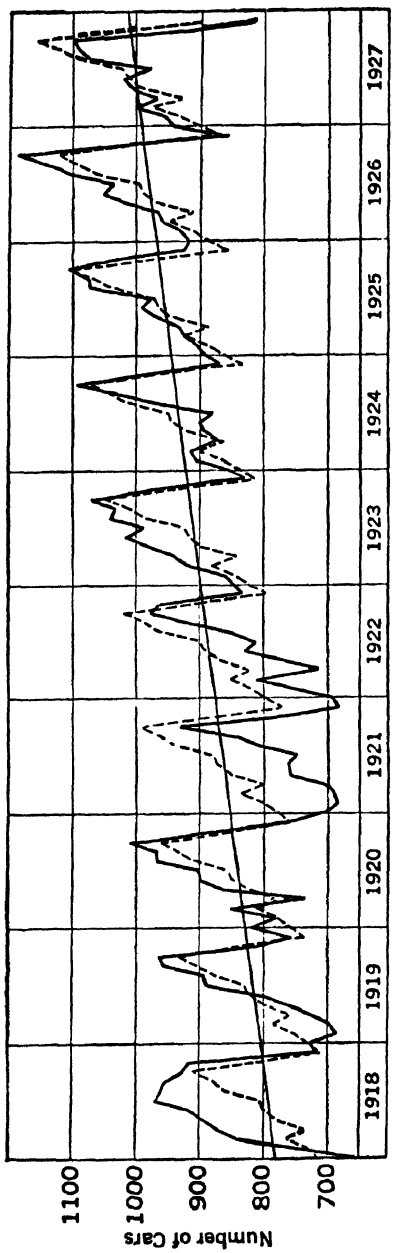


FIG. 63. — Freight Car Loadings in the United States, 1918–1927, with Line of Trend and Seasonal Pattern

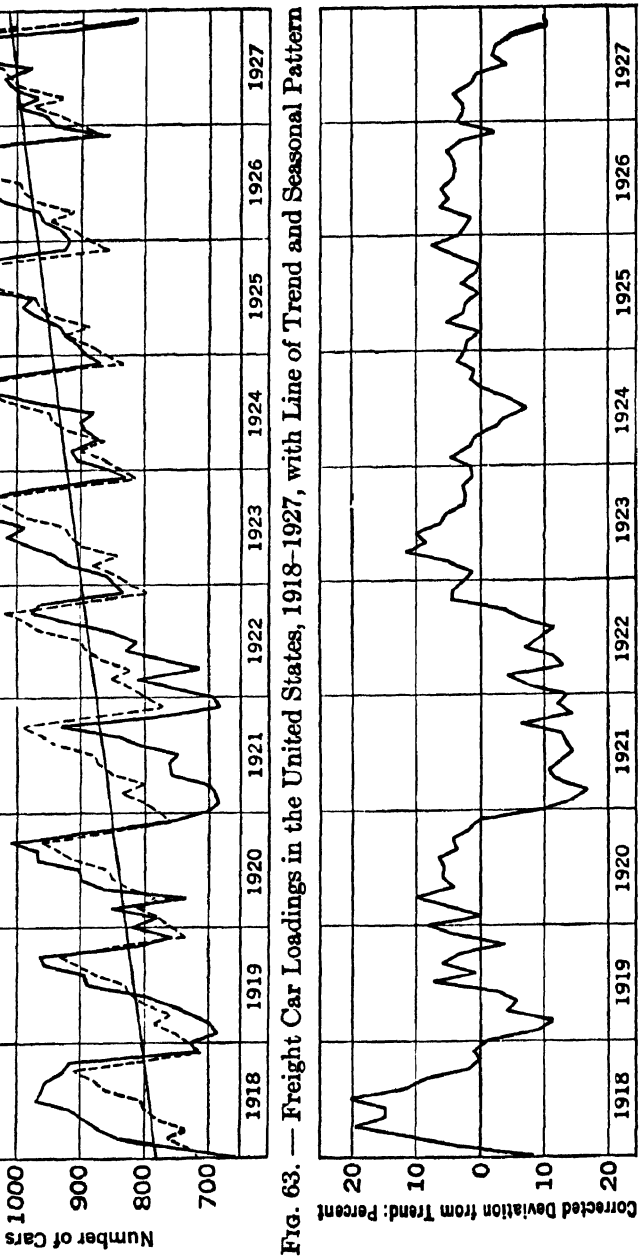


FIG. 64. — Cyclical and Accidental Fluctuations in Freight Car Loadings in the United States, 1918–1927

The process of analysis has now been completed. We have, for the given series, the equation to the line of secular trend, and from this the normal or trend value at any given date may be computed. The seasonal variations have been measured, and indices of these variations computed. Finally, the cyclical fluctuations (plus the unmeasurable random and accidental changes) have been isolated. These measurements of cyclical fluctuations may be used in studying the sequences of change in different economic series during business cycles, in comparing economic series in respect of the amplitude or duration of their cyclical movements, and in various other ways in the analysis of business cycles and the planning of business operations. Some of these applications are discussed in later sections.

General Considerations

Certain considerations not specifically mentioned above should be borne in mind in subjecting time series to the type of analysis described in this chapter. It is essential that the data employed be homogeneous, as regards sources, methods of quotation, coverage, etc. In addition, homogeneity in the conditions underlying the behavior of the particular series which are the objects of study is assumed. Homogeneity, as the term is here used, may not be defined in absolute terms. New factors are constantly being interjected into economic and social life. Homogeneity cannot be taken to mean static conditions. Yet the change must be orderly and, as regards major movements, reasonably continuous if the kind of analysis here discussed is to yield results. Abrupt dislocations that suddenly alter prevailing trends and existing seasonal patterns break the necessary homogeneity of statistical series. If the forces that caused these dislocations persist, and operate in orderly fashion, we mark a break in our series and subject the new period to analysis in its turn.

For the determination of a line of trend and the calcula-

tion of indices of seasonal variation, data extending over as long a period as possible should be employed (subject to the preceding qualification regarding fundamental discontinuities). Ten years may be suggested as a minimum period, though a much longer term of years is desirable. If interest attaches to cycles of long duration, rather than to the short-period business cycles with which the preceding account is concerned, our concept of trend, as well as that of cycles, must be modified. The minimum time period suitable for study must be correspondingly lengthened.

If a relatively short term of years is employed in the determination of trend, it is important that the terminal years be neither exceptionally high nor exceptionally low, as a result of cyclical or accidental movements. In general, the cyclical movements in the terminal years should be in "symmetrical phases," in Crum's phrase. Thus a cyclical rise at the beginning of the period should be balanced by a cyclical decline at the end.

It is logically improper to make correction for assumed seasonal movements in a time series unless the existence of true seasonal variations has been established. That is, a test should be applied to determine whether the observed departures of the various monthly indices from their average value (100) are attributable to the play of chance, or whether a true seasonal pattern is present. The basis of such tests of significance is discussed in Chapters XIV and XVIII, and a method appropriate to the present problem is developed in Chapter XV.

In fitting a line of trend, computing indices of seasonal variation and deriving, finally, a set of residual figures which are taken to reflect the cyclical fluctuations in an economic series we are, of course, abstracting from reality. As in all such abstractions, caution is necessary. Assumptions implicit in the various steps taken are likely to be forgotten. Thus the "cycles" plotted as deviations in Fig. 64 are distorted not only by the random and irregular fluctua-

tions to which attention has already been called. To the extent that the trend is inadequately or inaccurately defined by the particular function used, residual errors are present in the deviations. To the extent that seasonal movements are inaccurately measured by the seasonal indices employed, other residual errors are present. And if the trend is projected beyond the period covered by the fitting process, or if seasonal indices are used for periods not included in their calculation, new sources of possible error are introduced. The "cycles" that appear so definite and clear-cut in our tables may contain more than traces of many non-cyclical elements. It is often desirable to employ methods of analysis that carry us far from the original observations, but the dangers of misinterpretation and error are multiplied as we abstract from the reality of economic processes and business operations.

The methods of time series analysis described in this and the preceding chapter are adapted to a variety of economic and business purposes. But they do not constitute the only means of attack, in dealing with series ordered in time. Special problems may necessitate the use of somewhat more elaborate procedures.[1] For some purposes simpler methods will suffice. For other purposes it may be invalid to attempt to isolate and measure separately the influence of secular, seasonal, and cyclical forces. Economic science has yet to determine the precise nature of the interrelations among these categories of forces. In the light of this fact the discerning statistician will adapt his methods to the requirements of individual problems, as they arise.

REFERENCES

Chaddock, R. E., *Principles and Methods of Statistics*, Chap. 13.

Croxton, F. E. and Cowden, D. J., *Practical Business Statistics*, Chaps. 14, 16.

[1] Cf. the interesting technique employed by Wesley C. Mitchell in his monograph on the measurement of cyclical movements, to be published by the National Bureau of Economic Research. Preliminary drafts of Chapters I, II, and III have been released in mimeographed form by the National Bureau.

Crum, W. L. and Patton, A. C., *An Introduction to the Methods of Economic Statistics*, Chaps. 21, 22.

Davies, G. R. and Crowder, W. F., *Methods of Statistical Analysis in the Social Sciences*, Chap. 7.

Davies, G. R. and Yoder, D., *Business Statistics*, Chap. 5.

Day, E. E., *Statistical Analysis*, Chaps. 18, 19.

Kuznets, Simon, *Seasonal Variations in Industry and Trade.*

Kuznets, Simon, "Time Series," *Encyclopædia of the Social Sciences*, Vol. 14.

Lovitt, W. V. and Holtzclaw, H. F., *Statistics*, Chap. 11.

Mitchell, W. C., *Business Cycles, The Problem and Its Setting*, Chap. 3.

Moore, H. L., *Economic Cycles: Their Law and Cause.*

Moore, H. L., *Generating Economic Cycles.*

Persons, W. M., "Indices of Business Conditions." *Review of Economic Statistics.* Prel. Vol. I, 1919.

Rhodes, E. C., *Elementary Statistical Methods*, Chap. 12.

Riggleman, J. R. and Frisbee, I. N., *Business Statistics*, Chaps. 12, 13.

Smith, J. G., *Elementary Statistics*, Chaps. 13, 14.

Spurr, William A., "A Graphic Method of Measuring Seasonal Variation," *Journal of the American Statistical Association*, June, 1937.

Waugh, A. E., *Elements of Statistical Method*, Chap. 8.

CHAPTER IX

INDEX NUMBERS OF PHYSICAL VOLUME

Comprehensive and accurate records of physical production are of central importance to business interests, to government, and to economists. The appraisal of the market and the intelligent planning of production programs require knowledge of past production trends and present conditions. The credit policies of banking authorities and monetary policies of federal agencies are determined in good part with reference to the physical volume of goods being produced and marketed. The phases of business cycles are probably traced with more accuracy by production movements than by changes in any other economic element. The directions in which the productive efforts of an economy are being exerted are defined by records of the output of goods of different classes, such as capital goods and consumption goods. Changes in the course of years in the true standard of living of a nation must be measured in terms of the aggregate of physical goods produced.

The last twenty years have witnessed notable enlargements of the scope and improvements in the accuracy of measurements of production in the United States. Efforts of federal agencies, private organizations, and trade associations have combined to provide materially better statistics of output in agriculture, mining, and manufacture. More recently records of the volume of trade have been broadened and made more accurate. There are important gaps still, particularly as regards the output of finished, highly fabricated goods not easily enumerated in units of constant quality. But the statistics we have provide a full and

reasonably accurate record of monthly and annual movements of production.

Here, again, we face the problem of combining series relating to individual commodities. For scattered data on the output of oats, coal, gasoline, pig iron, automobiles, etc. do not define the general changes in output that are of interest to persons concerned with the larger aspects of economic change. He who would study the course of general production encounters a problem much like that presented to the student of general price movements. If the general trend of production is to be determined, or if the cyclical or seasonal swings of production are to be studied, the mass of individual figures must be reduced to the form of a single index, the significance of which may be easily comprehended. The present chapter deals with methods appropriate to the construction of such indices.

Index Numbers of Production Unadjusted for Trend and Seasonal Movements

An immediate and obvious obstacle to the combination of measures of output for different industries arises from differences in the units employed. Since bushels, tons, and gallons may not be added directly, the simple aggregative type of index is ruled out. One method of overcoming this difficulty is to reduce to relative terms the several output series that are to be combined. A relative number measuring the output of petroleum in 1936 as a percentage of output in 1922 may be averaged with similar relatives for bituminous and anthracite coal. The average may be a simple one, or the relatives for the several commodities may be weighted in proportion to the importance of the commodities in question. This procedure was illustrated in detail in the opening pages of Chapter VI.

An alternative method is to employ an index of the weighted aggregative type, keeping prices constant as between two periods being compared. In 1917, according

to the computations of the Price Section of the War Industries Board, the total value of the output of 90 raw materials in the United States was 34,748 millions of dollars. This figure represents, of course, a value total of the type $\Sigma(q_{1917}p_{1917})$ where q_{1917} represents the quantity of a given raw material produced in 1917 and p_{1917} represents the average price of that commodity in 1917. In 1918 both quantities and prices were different. If, however, we obtain another value aggregate using 1918 quantities and 1917 prices we shall have a figure differing from that for 1917 only in respect of the quantity factor. For the 90 raw materials in question this total, which is represented by the expression $\Sigma(q_{1918}p_{1917})$ amounted to \$35,169,000,000. The totals for 1918 and 1917 are comparable, being both in dollar units. The difference between them measures the change in physical production between the two years. As an index of this change we have

$$I = \frac{\Sigma(q_{1918}p_{1917})}{\Sigma(q_{1917}p_{1917})} = \frac{\$35{,}169}{\$34{,}748} = 101.2$$

This index will be recognized as one of the aggregative types discussed in Chapter VI, except that the p's and the q's are interchanged. When information concerning both quantities produced and average per unit prices is available, these aggregative indices, or the "ideal" index which is a combination of two such aggregative measures, may be employed for the measurement of quantity changes as well as for price changes. The "ideal" index, when used for this purpose, takes the form

$$I = \sqrt{\frac{\Sigma(q_1p_0)}{\Sigma(q_0p_0)} \times \frac{\Sigma(q_1p_1)}{\Sigma(q_0p_1)}}$$

where q_0 and p_0 represent the quantities and prices of the individual commodities in the base year, while q_1 and p_1 represent quantities and prices in the given year. The procedure in the computation of such an index is identical

with that employed in computing the "ideal" price index, with prices and quantities reversed. This formula may be modified, as was the corresponding price index, to

$$\frac{\Sigma(p_0 + p_1)q_1}{\Sigma(p_0 + p_1)q_0}$$

or to a form in which the p's come from some intermediate year. In one form or another, the aggregative type of

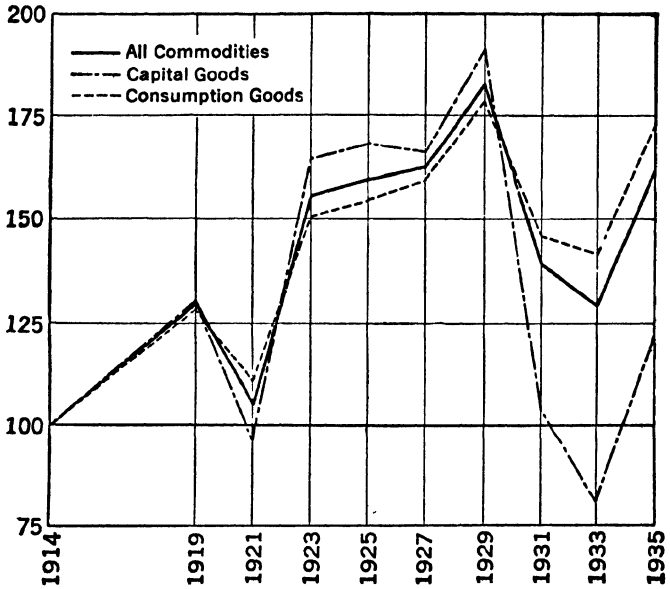


FIG. 65. — Changes in the Physical Volume of Manufacturing Production in the United States, 1914–1935. All Commodities, Capital Goods and Consumption Goods

index is well adapted to the requirements of an index of physical volume.[1]

The aggregative procedure lends itself readily to the con-

[1] Since the price or value factor enters in the derivation of such an index, whether it be constructed from relative numbers or from value aggregates, no quantity index is completely divorced from pecuniary measurements. For a discussion of this point, and of other logical problems involved in the construction of index numbers of production, see Arthur F. Burns, "The Measurement of the Physical Volume of Production," *Quarterly Journal of Economics*, February, 1930.

struction of index numbers for commodity groups. This is desirable in the study of production movements, as it is for prices. The significant features of production changes over a given period may be far more clearly revealed by measurements of relative changes in the output of different classes of goods than by a general index of production.

Changes in the volume of production of various classes of manufactured goods during the period 1914–1935 are indicated by the following measurements, constructed by the National Bureau of Economic Research. The basic data, which were compiled by the Census of Manufactures, provide the quantity and (by derivation) the unit price records required for the "ideal" formula. That formula, slightly modified for working purposes, was employed in the construction of these index numbers.

TABLE 81

Index Numbers of the Physical Volume of Production of Manufactured Goods in the United States, 1914–1935 [1]

	All industries	*Durable goods*	*Semi-durable goods*	*Perishable goods*	*Goods destined for human consumption*	*Goods destined for capital equipment and construction*
1914	100.0	100.0	100.0	100.0	100.0	100.0
1919	129.5	141.7	120.9	123 2	129.1	129.5
1921	104.5	99.6	104.6	108 9	109.4	91.8
1923	155.8	183.7	140.2	135.4	150.4	164.3
1925	159.5	185.2	141.8	144.4	154.0	167.7
1927	163.3	177.2	151.0	154.9	159.5	166.7
1929	183.7	210.9	162.5	170.9	177.7	192.0
1931	138.2	112.3	137.4	154.9	146.9	103.7
1933	128.0	91.4	140.1	144.4	142.6	81.3
1935	160.5	143.9	164.4	163.9	171.3	122.5

Selected measurements from Table 81 are shown graphically in Fig. 65.

[1] Constructed by the National Bureau of Economic Research, New York. See *Economic Tendencies in the United States* for a statement on procedure.

Adjusted Index Numbers of Production

In the analysis of time series we have seen that cyclical fluctuations are often the objects of primary interest. This is particularly true in the study of physical volume, for changes in the volume of production and trade are features of fundamental importance in business cycles. Methods have been explained, in the preceding chapters, by means of which we seek to measure the cyclical fluctuations in individual series (fluctuations which are inextricably entangled with accidental movements of major and minor degree). An obvious next step, in the study of general business conditions, is the combination of the cyclical-accidental movements in a number of series into a single index. The utility of such an index of changes in the physical volume of production in the course of the business cycle is evident.

When annual data are employed the construction of an index of these cyclical changes is simple. No problem of seasonal variation enters, and secular trend alone has to be taken account of. Two different methods by which this may be done present themselves. Edmund E. Day, a pioneer in this field of economic research, has tested both methods.

The first involves the fitting of an appropriate line of trend to each of the constituent series. The actual items are then expressed as percentages of the corresponding trend values. When this has been done for each series, the final adjusted index for a given year is obtained by taking a weighted average of these percentages for that year. Each commodity may be weighted in this averaging process, as in the calculation of the unadjusted index. The resulting adjusted index is in terms of relatives, but these relatives refer to a hypothetical "normal," instead of to any fixed base. This is the desired index of cyclical-accidental changes in the physical volume of production. With monthly data the process is the same, except that

before being averaged, the deviations from trend are corrected to eliminate the influence of recurrent seasonal movements.

In the process of averaging deviations from trend, account should be taken of the relative variability of the series being combined. As an example, we may consider the combination of data of pig iron production and cattle receipts in a general index of production. Reducing pig iron production to terms of "seasonably adjusted deviations from trend," we obtain a series marked by rather extreme fluctuations. The standard deviation of this adjusted series, for a given period, was 27. For cattle receipts, correspondingly adjusted, the standard deviation was 11. In any combination of the two series of percentage deviations the more widely fluctuating pig iron measurements will exercise a dominant influence, unless correction is made. The use of weights defining the relative economic importance of the two series will not prevent distortion due to the greater variability of the pig iron series.

One way out of the difficulty is to divide the deviations from trend by the respective standard deviations, before averaging. This gives an index in standard deviation units. Another procedure involves the combination of the "economic weight" and the standard deviation of each series in a weighting factor to be applied directly to the percentage deviations from trend. The economic weight is divided by the corresponding standard deviation, in making the combination. The method is illustrated below.

Series	*Economic weight*	*Standard deviation*	*Economic weight ÷ standard deviation*
Pig iron production	20	27	.747
Cattle receipts	4	11	.363

The final weighting factors are the figures given in the last column. These may, of course, be rounded off when a number of series are to be averaged.[1]

[1] This useful method of combining economic weights and standard devia

The alternative method of combining economic series is simpler. Each unadjusted index possesses a trend which is "a composite of the persistent tendencies of the several original series upon which the unadjusted index is based." It is possible to measure this trend, instead of the separate original trends, and secure the adjusted index directly from the unadjusted. Day's results indicate that there is no loss of accuracy in the use of the simpler method.

AN INDEX OF INDUSTRIAL ACTIVITY

This procedure, with certain modifications, is well exemplified in an "Index of Industrial Activity in the United States," constructed by the Chief Statistician's Division of the American Telephone and Telegraph Company.[1] The elements of this index are monthly data; seasonal corrections are therefore necessary. When these corrections have been made a general index measuring long-term growth and cyclical-accidental fluctuations, in combination, is constructed by averaging 11 series, with appropriate weights.[2] This index is shown for the period 1899–1937, with line of trend, in Fig. 66. The trend line was fitted by least squares to data for the period 1899–1930, with the war years, 1917–1918, omitted.

When each monthly value of the index is expressed as a percentage deviation from the corresponding trend value, the measurements presented in Table 82, and graphically portrayed in Fig. 67, are obtained. The cyclical-accidental

tions has been employed by G. W. Starr, Director of the Bureau of Business Research of Indiana University. I am indebted to him for the example.

[1] This index has been constructed for the use of the staffs of the Bell system companies, and is not available for distribution. It is published here by courtesy of the American Telephone and Telegraph Company.

[2] The following series were used for the later years of the period covered: steel ingot production, pig iron production, automobile passenger car production, building contracts awarded (on a square foot basis), cotton consumption, wool consumption, slaughter of cattle and hogs, newsprint consumption, miscellaneous freight car loadings, electric power consumption, and employment in manufacturing industries. Since employment is included, the index goes slightly beyond the field of strict physical production. It is intended to be an index of industrial activity.

FIGURE 66

THE GROWTH OF INDUSTRIAL ACTIVITY IN THE UNITED STATES

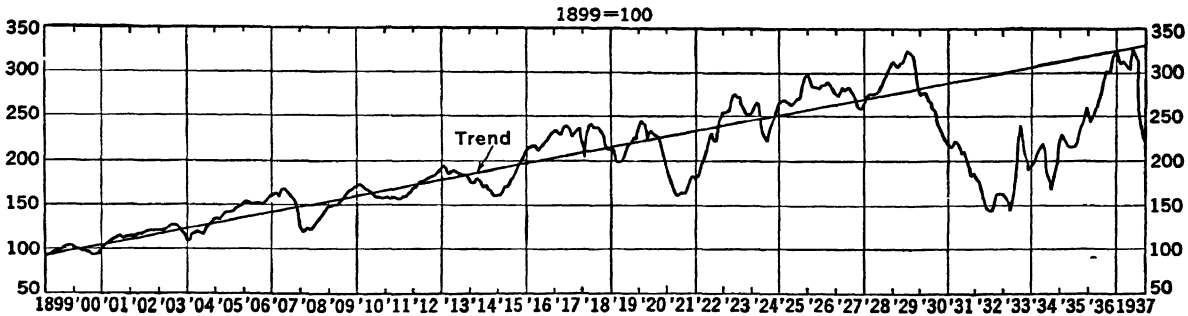


FIGURE 67

INDUSTRIAL ACTIVITY AS RELATED TO LONG-TERM GROWTH

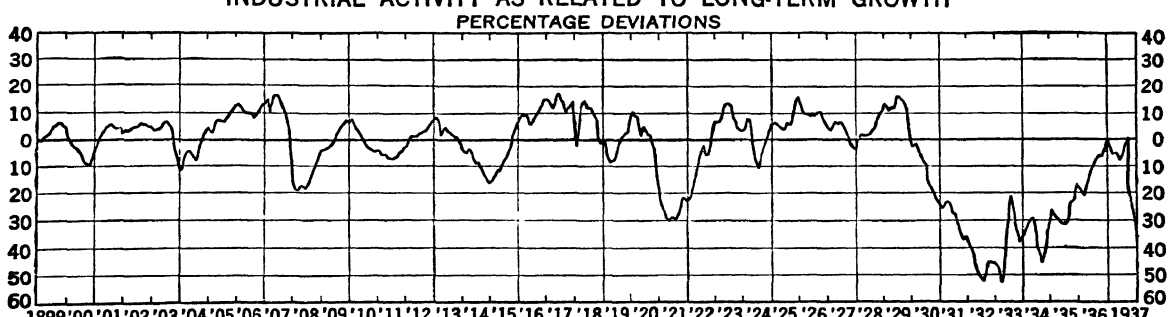


TABLE 82

Industrial Activity as Related to Long-Term Growth, 1899–1937

(Percentage deviations)

	1899	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909	1910	1911
January	+ 0.2	+ 0.5	− 2.2	+ 3.4	+ 3.0	− 11.3	+ 3.5	+ 13.1	+ 13.7	− 18.9	− 4.1	+ 7.6	− 4.0
February	− 0.7	− 0.1	+ 0.3	+ 2.8	+ 3.0	− 6.4	+ 2.4	+ 12.7	+ 15.3	− 19.0	− 3.6	+ 6.4	− 5.9
March	+ 0.9	− 2.0	+ .4	+ 3.2	+ 3.7	− 4.9	+ 5.8	+ 10.9	+ 10.0	− 17.0	− 3.3	+ 4.5	− 6.0
April	+ 1.4	− 2.9	+ 3.3	+ 4.6	+ 4.1	− 3.8	+ 7.3	+ 9.9	+ 16.1	− 17.7	− 2.4	+ 3.6	− 5.4
May	+ 1.9	− 3.2	+ 4.6	+ 4.6	+ 5.5	− 4.7	+ 6.9	+ 9.6	+ 16.3	− 18.3	− 1.3	+ 2.4	− 7.1
June	+ 2.7	− 4.3	+ 5.1	+ 4.5	+ 6.4	− 6.5	+ 6.7	+ 9.8	+ 16.4	− 17.1	+ 0.1	+ 0.3	− 7.2
July	+ 4.8	− 6.3	+ 5.8	+ 5.1	+ 6.1	− 8.1	+ 6.3	+ 10.0	+ 14.6	− 14.6	+ 2.3	− 1.8	− 7.4
August	+ 5.3	− 8.6	+ 4.4	+ 6.2	+ 5.0	− 6.0	+ 7.5	+ 7.8	+ 12.4	− 12.2	+ 3.7	− 3.0	− 6.8
September	+ 6.4	− 9.1	+ 3.7	+ 5.3	+ 3.1	− 1.4	+ 9.0	+ 8.8	+ 9.5	− 11.3	+ 4.8	− 3.3	− 6.6
October	+ 6.4	− 9.6	+ 3.6	+ 5.6	− 0.5	+ 0.1	+ 9.8	+ 10.3	+ 6.0	− 9.4	+ 5.2	− 4.3	− 5.3
November	+ 5.1	− 8.7	+ 4.2	+ 4.5	− 6.4	+ 2.8	+ 11.2	+ 11.4	− 3.6	− 6.4	+ 7.1	− 4.9	− 4.7
December	+ 4.4	− 4.5	+ 1.7	+ 4.7	− 11.1	+ 4.1	+ 12.6	+ 13.6	− 16.2	− 4.2	+ 6.1	− 4.4	− 2.7
Average	+ 3.2	− 4.9	+ 3.0	+ 4.6	+ 1.8	− 3.8	+ 7.4	+ 10.7	+ 9.2	− 13.8	+ 1.2	+ 0.3	− 5.8

	1912	1913	1914	1915	1916	1917	1918	1919	1920	1921	1922	1923	1924
January	− 2.6	+ 8.4	− 5 0	− 15.6	+ 8.6	+ 14.9	− 2.8	− 1.6	+ 10.7	− 24.0	− 22.2	+ 5.9	+ 5.1
February	+ 0.5	+ 7.3	− 5.8	− 13.8	+ 8 7	+ 13.3	+ 5.9	− 6.8	+ 8.4	− 26.3	− 19.5	+ 6.6	+ 7.8
March	+ 1.3	+ 1.6	− 3.7	− 11.6	+ 9.5	+ 11.5	+ 12.8	− 8.7	+ 8.3	− 29.8	− 16.2	+ 8.6	+ 6.9
April	+ 0.7	+ 3.4	− 4.8	− 12.1	+ 8 9	+ 15.3	+ 14.5	− 8.0	+ 0.5	− 30.2	− 12.0	+ 12.9	+ 2.4
May	+ 1.3	+ 4.8	− 7.9	− 10.1	+ 5.6	+ 17.1	+ 11.8	− 7.3	+ 4.6	− 29.0	− 8.6	+ 13.1	− 5.7
June	+ 2.4	+ 3.0	− 9.0	− 8.1	+ 6.9	+ 15.3	+ 11.8	− 4.9	+ 3.1	− 29.1	− 3.6	+ 12.1	− 10.0
July	+ 2.4	+ 2.0	− 8.4	− 7.5	+ 7.5	+ 13 9	+ 11.8	− 0.7	+ 1.5	− 30.0	− 1.9	+ 12.5	− 10.8
August	+ 2.7	+ 1.6	− 11.1	− 5 5	+ 9 2	+ 9.4	+ 9 5	+ 0.8	+ 1.2	− 28.0	− 6.4	+ 7.9	− 6.0
September	+ 3.1	+ 0.8	− 12.5	− 2.9	+ 10.7	+ 11.4	+ 7.6	+ 1.6	− 1.7	− 25.7	− 5.8	+ 6.4	− 3.1
October	+ 4.3	+ 0.7	− 14.4	+ 1.1	+ 12.6	+ 12 1	− 0.3	+ 2.8	− 5.5	− 21.9	− 1.3	+ 3.6	− 1.5
November	+ 5.9	− 0.9	− 16.4	+ 3.1	+ 14.6	+ 14 0	− 1.5	+ 2.0	− 12 5	− 22.1	+ 3.4	+ 2.9	+ 1.6
December	+ 6.7	− 4.6	− 16.5	+ 6.4	+ 15.0	+ 7.6	0.0	+ 8.4	− 17.6	− 23.1	+ 6.5	+ 3.0	+ 5.1
Average	+ 2.4	+ 2.3	− 9.6	− 6 4	+ 9 8	+ 13 0	+ 6.8	− 1.9	+ 0.1	− 26.6	− 7.3	+ 8.0	− 0.7

	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937
January	+ 6.4	+ 12.1	+ 3.3	+ 0.7	+ 12 0	− 1.7	− 25.5	− 39.2	− 47.1	− 35.2	− 26.6	− 20.3	− 3.4
February	+ 6.1	+ 9.8	+ 4.5	+ 2.0	+ 10.5	− 1.5	− 24.2	− 40.8	− 48.4	− 32.3	− 27.4	− 23.2	− 5.0
March	+ 5 2	+ 9.3	+ 6.4	+ 1.4	+ 11.9	− 5.1	− 22.8	− 44.0	− 52.6	− 30.2	− 28.6	− 20.8	− 3.5
April	+ 4.6	+ 9 1	+ 5.6	+ 1.3	+ 11.7	− 5.8	− 23.5	− 48.6	− 47.7	− 29.8	− 30.8	− 19 0	− 5.6
May	+ 3.4	+ 8.4	+ 6.4	+ 1.3	+ 13 3	− 7.4	− 25.3	− 50.9	− 39.5	− 29.3	− 31.7	− 18.0	− 6.0
June	+ 4.0	+ 9.1	+ 5.4	+ 1.8	+ 16.4	− 10.1	− 28.3	− 52.0	− 28.5	− 30.9	− 31.6	− 14.1	− 7.7
July	+ 5.9	+ 8.8	+ 3.9	+ 3.6	+ 15.7	− 15.1	− 28.3	− 52.9	− 21.6	− 38.9	− 30.9	− 10.9	− 5.3
August	+ 5.5	+ 10.2	+ 2.3	+ 4.5	+ 15.0	− 16.9	− 30.7	− 50.4	− 25.3	− 40.4	− 25.8	− 7.0	− 0 8
September	+ 6.5	+ 10.4	− 0.3	+ 7.2	+ 14.0	− 18.5	− 34.0	− 45.9	− 30.1	− 44.9	− 24.2	− 6.1	− 5.2
October	+ 10.6	+ 9.0	− 2.6	+ 9.7	+ 11.5	− 21.2	− 37.1	− 45.8	− 32.8	− 39.8	− 21.1	− 6.9	− 12.6
November	+ 14.8	+ 6.7	− 3.8	+ 11.1	+ 2.6	− 22.3	− 36.1	− 45.6	− 37.7	− 38.0	− 18.5	− 3.0	− 25.5
December	+ 16.1	+ 4.5	− 3.4	+ 13.7	− 2.8	− 23.7	− 36 3	− 45.7	− 36.3	− 31.8	− 15.3	+ 0.7	− 32.3
Average	+ 7.4	+ 9.0	+ 2.3	+ 4.9	+ 11.0	− 12.4	− 29.3	− 46.8	− 37.3	− 35.1	− 26.0	− 12.4	− 9.4

fluctuations in industrial activity, as represented by the 11 series employed, are traced by the movements of this index.

Index of Industrial Production of the Board of Governors of the Federal Reserve System

A comprehensive monthly index of production in mining and manufacturing industries of the United States is constructed by the Division of Research and Statistics of the Board of Governors of the Federal Reserve System. This index is designed to serve current needs. In the selection of its components emphasis has been placed upon the promptness with which basic data on the output of industrial commodities become available, as well as upon their accuracy and representativeness.

The chief points of general interest relating to this index may be briefly noted.

Coverage. The index is derived from 60 individual series, measuring production in some 35 industries. Approximately 80 per cent of the total industrial production of the United States is represented directly or indirectly in the index.

Base period. The base of the published relatives is daily average production during the three years 1923, 1924, and 1925. The final indices appear as relatives on this base, both with and without seasonal correction.

Character of data used. For each commodity production is computed in terms of average output per working day. By this method distortion due to changes from one month to the next in the number of Sundays and holidays included is avoided.

Form of index number. The index is of the weighted aggregative type. Original quantity figures are multiplied by weighting factors which convert them into common units (i.e., values, in dollars). In deriving the final index, the aggregate for a given date is expressed as a percentage of the base-period aggregate.

Weighting factors. For mineral products the weight for each commodity is its average per unit value in the base period. For manufactured products the weight for a given commodity is the per unit "value added" (i.e., added by manufacture), modified to the extent that the commodity in question is taken to represent other manufactured products not directly included in the index. These "weights" thus correspond to p's in the aggregative formula $\frac{\Sigma(q_1p_0)}{\Sigma(q_0p_0)}$, except that for a manufactured product the p is a "price"

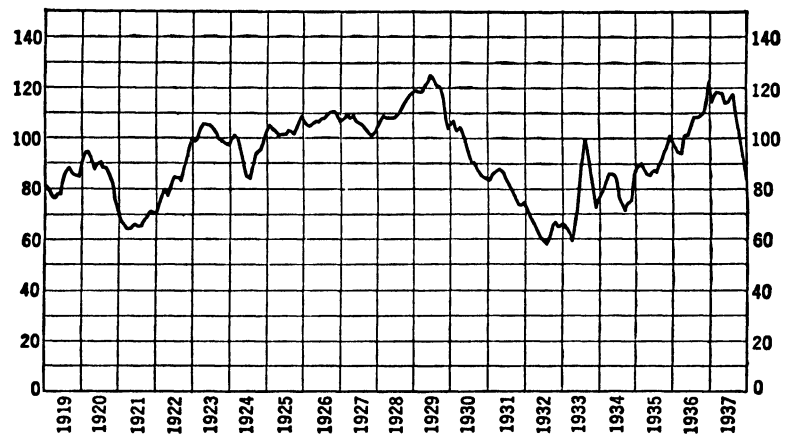


FIG. 68. — Physical Volume of Industrial Production in the United States, 1919–1937 (1923–1925 average = 100)

for the services of agents of fabrication, with a modification to allow the given commodity to represent similar products for which quantity data are not available. The weights for manufactured goods were drawn from the Census of Manufactures for 1923. The p_0 used to weight the q for manufactures is thus not strictly a base-period price.[1]

Adjustment for seasonal variation. No correction for trend is made, but in one form of the index an adjustment is made to eliminate the effect of seasonal fluctuations in the

[1] Weighting factors were modified for the period 1919–1922 by the combination of weights for 1919 with those for the base period.

TABLE 83

Index of Industrial Production, Board of Governors of the Federal Reserve System, 1919–1937

(Adjusted for seasonal variation. 1923–1925 average = 100)

Month	1919	1920	1921	1922	1923	1924	1925	1926	1927	1928
Jan.	82	95	67	73	99	100	105	106	107	107
Feb.	79	95	66	76	100	102	104	105	108	109
March	76	93	64	80	103	100	103	106	110	108
April	78	88	64	77	106	95	102	107	108	108
May	78	90	66	81	106	89	102	106	109	108
June	83	91	65	85	106	85	102	108	107	108
July	87	89	65	85	104	84	103	108	106	109
Aug.	89	89	67	83	103	89	103	110	106	110
Sept.	87	86	68	88	100	94	101	111	104	113
Oct.	86	83	71	93	99	95	104	111	102	115
Nov.	85	76	71	97	98	97	107	110	101	117
Dec.	86	72	70	100	97	101	109	107	102	118
Annual index	83	87	67	85	101	95	104	108	106	111

Month	1929	1930	1931	1932	1933	1934	1935	1936	1937
Jan.	119	106	83	72	65	78	90	97	114
Feb.	118	107	86	69	63	81	89	94	116
March	118	103	87	67	59	84	88	93	118
April	121	104	88	63	66	85	86	101	118
May	122	102	87	60	78	86	85	101	118
June	125	98	83	59	91	83	87	104	114
July	124	93	82	58	100	76	86	108	114
Aug.	121	90	78	60	91	73	88	108	117
Sept.	121	90	76	66	84	71	91	109	111
Oct.	118	88	73	67	76	73	95	110	102
Nov.	110	86	73	65	72	74	96	114	88
Dec.	103	84	74	66	75	86	101	121	84
Annual index	119	96	81	64	76	79	90	105	110

output of individual commodities. Seasonal indices were computed by averaging the ratios of actual data to twelve-month moving averages. (See Chapter VIII.) Where there was evidence of progressive change in the seasonal pattern, the seasonal adjustments for a given commodity were modified from year to year. The actual adjustment for

seasonal change is made by dividing the daily average output of a given commodity in a stated month by the seasonal index for that month, expressed as a ratio (i.e., as 1.10, if the conventional index were 110). The seasonally adjusted q would thus be reduced if the seasonal index were above 1.00, raised if the seasonal index were below 1.00. In the construction of the seasonally corrected index, these adjusted q's are used in the aggregative formula previously described.[1]

Monthly values of this index are given in Table 83, for the period 1919–1937. The index is shown graphically in Fig. 68 on page 317.

Derived Indices of Production and Productivity

It is possible, where suitable records of value of product and indices of price changes are available, to derive an index of production by indirection. In the case of a single commodity it is obvious that $pq \div p = q$. (Here q represents the number of physical units produced, p represents average per unit price, and pq is the aggregate value.) A similar process is possible in handling statistics relating to a number of commodities, in combination. Indeed, the records may be in the form of relatives, or index numbers, covering a number of months or years. Division of a value index by a price index relating to the commodities included in the value index will yield an index measuring changes in physical output.

This procedure may sometimes be used to obtain measurements that could not possibly be built up by combining a number of individual records. Whether the method is applicable in a given instance depends upon the comparability of the price and value index numbers. The strict

[1] A detailed description of the constituents of this index and of the procedure employed in its construction is given in the *Federal Reserve Bulletin* for February, 1927. Revisions are noted in the issues of that *Bulletin* for March, 1932, Sept., 1933, Nov., 1936, and March, 1937. The index appears in current issues of the *Federal Reserve Bulletin*.

requirement that the price index relate to precisely the commodities included in the value index cannot generally be met. If we assume that a given price index is fairly representative of the commodities covered by the value records, and if the formula employed in the construction of the index is appropriate, the method may be justified as a means of approximating the required index of physical output.

An example of such a procedure is furnished by the materials in Table 84. These illustrate a method used in deriving an index of production of manufactured goods. The indices in col. (3) are derived directly from the aggregate figures on "value added by manufacture." The indices in col. (4) measure changes in average "value added" per unit, or cost of fabrication per unit, of manufactured goods. (This is, in effect, a price index, the price covering the services of manufacturing agents in the process of fabrication.) This series of index numbers is based upon records available for a representative sample of manufacturing industries. The general index of manufacturing production,

TABLE 84

Illustrating the Derivation of Index Numbers of the Physical Volume of Manufacturing Production, 1923–1929 [1]

(1)	(2)	(3)	(4)	(5)
Year	*Total value added, all census industries* (*in millions of dollars*)	(*in relatives*)	*Index of value added per unit of product, industries included in sample*	*Derived index of physical volume of production* (3) ÷ (4)
1923	25,850	100.0	100.0	100.0
1925	26,778	103.6	97.3	106.4
1927	27,585	106.7	92.4	115.4
1929	31,844	123.2	96.8	127.3

[1] This table is taken from *Economic Tendencies in the United States*, New York, National Bureau of Economic Research, 308.

relating to all industries, is derived by dividing the relatives for total "value added" by the index numbers measuring changes in "value added" per unit of product (with a suitable shift in the decimal point).

The derived measurements given in col. (5) of Table 84 are probably more accurate than index numbers based upon directly enumerated physical products. For the gaps in the coverage of the latter are serious. Limitations of coverage are the more serious in that the excluded industries are in many cases just the new, rapidly developing industries the output of which is growing most rapidly.

A somewhat similar process of derivation is employed in the construction of measurements of industrial productivity. It is impossible, by direct observation, to compile records of output per man or per man-hour over any considerable area of industrial activity. However, given accurate indices of physical production and comparable records of number of workers employed or of man-hours worked, one may derive index numbers measuring changes in productivity.

An example of this procedure is given in Table 85. The measurements given should be regarded as approximations only.

TABLE 85

Index Numbers of Physical Volume of Production, Man-Hours Worked and Output per Man-Hour, Manufacturing Industries of the United States, 1929–1935

Year	*Physical volume of manufacturing production*	*Total number of man-hours worked*	*Estimated output per man-hour*
1929	100	100	100
1931	75	66	114
1933	70	60	117
1935	87	70	124

Between 1929 and 1935 the total volume of manufacturing production declined 13 per cent. The number of man-

hours worked decreased by 30 per cent, however. The indicated gain in output per man-hour was 24 per cent.

Measurements such as these are of unquestioned value to the student of industrial change, but their limitations should be clearly stated. The initial necessity of full comparability between the output and employment records has been mentioned. Discrepancies here may lead to serious errors in the derived measurements. More difficult to detect are technical industrial changes that do not appear in the statistical records. Changes in the quality of the goods represented in the production index may lessen the accuracy of that index, and affect the productivity measurements. If employment is measured in terms of number of men employed, the resulting index of per capita output may be seriously distorted by changes in the length of the working week. Again, if only direct labor is enumerated in the employment index, a shift in technical methods that involves the use of a much larger proportion of indirect labor may lead to a great advance in apparent productivity, which far exceeds the real gain. Some of the gain that apparently follows the increased mechanization of a plant or a process is of this fictitious sort. Labor that precedes the direct act of production, and servicing and supervising labor, may have replaced direct labor. Failure to take account of the contributions of these indirect applications of labor may lead to grossly exaggerated measures of productivity gains.

The purpose of the preceding pages has been to exemplify procedures used in the measurement of changes in production, with incidental reference to related problems. While there is no one standard method, it will be clear that the construction of quantity index numbers requires no involved procedure. Certain special problems — of weighting, of measuring secular and seasonal movements, of ensuring comparability when methods of derivation are employed — have been noted. In addition, most of the problems that

bulk large in the construction of price index numbers are faced in this area also. The task of obtaining accurate, homogeneous series of basic data entails no less careful field work in production than in prices. Quality changes lessen the accuracy of both types of index numbers. Comparisons over considerable time periods are rendered inaccurate by such quality changes and perhaps even more by changes in "regimen"— in the complex of tastes, consuming habits, and technical methods that determines the weighting factors used in the construction of index numbers. In spite of these difficulties substantial progress has been made in recent years in the improvement of measures of industrial activity of the type discussed in this chapter. More comprehensive and more accurate data are being compiled, and technical standards in the construction of index numbers are being steadily raised. These gains are contributing to a notable advance in our knowledge of economic processes.

REFERENCES

Bliss, Charles A., "*Production in Depression and Recovery,*" Bulletin 58, National Bureau of Economic Research, Nov. 15, 1935.

Burns, Arthur F., "*The Measurement of the Physical Volume of Production,*" Quarterly Journal of Economics, Feb., 1930.

Burns, Arthur F., *Production Trends in the United States Since 1870.*

Day, E. E., "An Index of the Physical Volume of Production." *Review of Economic Statistics,* Sept. 1920, Jan. 1921.

Day, E. E. and Thomas, W., *The Growth of Manufactures,* 1899–1923.

Leong, Y. S., "Indexes of the Physical Volume of Production of Producers' Goods, Consumers' Goods, Durable Goods and Transient Goods." *Journal of the American Statistical Association,* June, 1935.

Mills, F. C., *Economic Tendencies in the United States,* Appendices 1, 4.

Mitchell, W. C., *Business Cycles, The Problem and Its Setting,* Chap. 3.

Perry, F. G. and Silverman, A. G., "A New Index of the Physical

Volume of Canadian Business," *Journal of the American Statistical Association*, June, 1929.

Persons, W. M., *Forecasting Business Cycles.*

Snyder, Carl, *Business Cycles and Business Measurements*, Chaps. 2–5.

Weintraub, David, "Unemployment and Increasing Productivity." In *Technological Trends and National Policy*, National Resources Committee, June, 1937 (75th Congress, 1st Session, House Document No. 360).

CHAPTER X

THE MEASUREMENT OF RELATIONSHIP: LINEAR CORRELATION

In discussing averages and measures of dispersion and skewness we have been dealing with methods of describing a single frequency distribution. The arrangement of the values of a single variable along a scale may be portrayed by means of these measures, which enable the central value to be determined and the character of the distribution about that central value to be described. In the analysis of time series a somewhat different problem has been faced. In such cases we are concerned with the changing values of a variable factor with the passage of time, and seek to determine the degree to which the changes in value are due to the play of different forces — the secular trend and cyclical, seasonal, and accidental factors. The preceding chapters dealt with methods by which we might measure the effect upon a given series of each of these factors (with the exception of accidental fluctuations).

Certain of these methods are applicable to the problem now before us. It was found that in dealing with time series the relationship between time and the long term trend factor may be described by a definite mathematical equation. That is, trend or growth seems to be a function of time for many economic series. Where such a relationship prevails, whether it hold precisely or only approximately, there is a distinct advantage in securing a mathematical expression which describes it. A similar but much broader problem is now to be discussed. If it is possible in dealing with time series to secure a definite mathematical equation for the relation between time and the normal values of the

items in a given series, cannot the same device be employed in studying the relationship between other variables? Can we not define, mathematically, the relation between cotton production and the price of cotton, between corn yield and rainfall, between earnings and the output of labor? If this can be done, it will place in the hands of the economist a very powerful tool, giving his methods something of the precision which attaches to the work of the physical scientist.

The Relation between Number of Taxable Personal Incomes and Motor Vehicle Registration

As a typical problem we may consider the relation between the number of taxable personal incomes and the number of passenger automobiles registered, by states in 1934. The figures are given in columns (2) and (3) of Table 86.[1]

These figures are plotted in Fig. 69, each dot representing the relation between the number of taxable incomes and the number of registered passenger cars for a given state. Such a figure is termed a "scatter diagram." It is clear from this diagram that there is a relationship between the two variables. In general, the states with a large number of taxable personal incomes are also those having a large number of motor vehicle registrations. The relationship, however, is not perfect. Two states with the same number of taxable incomes may differ quite widely in the number of registered vehicles. Thus both Rhode Island and Colorado

[1] Nine states for each of which there were more than 100,000 individual income tax returns and more than 685,000 passenger cars registered in 1934 have not been included. The observations for these states, some of which are materially affected by the presence of important industrial centers, depart rather widely from those for the remaining states, and are marked by a functional relationship between personal incomes and motor vehicle ownership somewhat different from that prevailing for the country at large. The states thus excluded are New York, Pennsylvania, New Jersey, Illinois, Massachusetts, Michigan, Texas, Ohio, and California. The state of Washington has also been excluded, since the income tax returns for that state are combined with those of Alaska, in the reports of the Bureau of Internal Revenue. The results are to be interpreted, of course, with these restrictions in mind.

Table 86

Taxable Personal Incomes and Passenger Automobile Registration in Thirty-Eight States, 1934

(1) State	(2) No. of taxable personal incomes, 1934 (thousands) X	(3) No. of passenger cars registered, 1934 (thousands) Y	(4) XY	(5) X^2	(6) Y^2
Alabama	23	192	4,416	529	36,864
Arizona	11	80	880	121	6,400
Arkansas	13	162	2,106	169	26,244
Colorado	31	246	7,626	961	60,516
Connecticut	91	310	28,210	8,281	96,100
Delaware	11	45	495	121	2,025
Florida	33	280	9,240	1,089	78,400
Georgia	38	317	12,046	1,444	100,489
Idaho	9	91	819	81	8,281
Indiana	70	680	47,600	4,900	462,400
Iowa	48	591	28,368	2,304	349,281
Kansas	36	453	16,308	1,296	205,209
Kentucky	35	295	10,325	1,225	87,025
Louisiana	37	199	7,363	1,369	39,601
Maine	21	141	2,961	441	19,881
Maryland	84	288	24,192	7,056	82,944
Minnesota	67	594	39,798	4,489	352,836
Mississippi	13	141	1,833	169	19,881
Missouri	98	632	61,936	9,604	399,424
Montana	17	97	1,649	289	9,409
Nebraska	27	350	9,450	729	122,500
Nevada	5	26	130	25	676
New Hampshire	17	91	1,547	289	8,281
New Mexico	8	67	536	64	4,489
N. Carolina	32	385	12,320	1,024	148,225
N. Dakota	10	130	1,300	100	16,900
Oklahoma	39	403	15,717	1,521	162,409
Oregon	27	233	6,291	729	54,289
Rhode Island	31	124	3,844	961	15,376
S. Carolina	15	182	2,730	225	33,124
S. Dakota	8	146	1,168	64	21,316
Tennessee	38	299	11,362	1,444	89,401
Utah	11	85	935	121	7,225
Vermont	10	69	690	100	4,761
Virginia	48	317	15,216	2,304	100,489
W. Virginia	30	167	5,010	900	27,889
Wisconsin	93	589	54,777	8,649	346,921
Wyoming	7	52	364	49	2,704
Totals	1,242	9,549	451,558	65,236	3,610,185

had 31,000 taxable personal incomes in 1921, yet the former had 124,000 passenger cars registered, while the latter had 246,000. Were the relationship perfect a single and unchanging value of the Y-variable would always be found paired with a given value of the X-variable.

Our first problem is the derivation of an equation to describe this relationship which, while not perfect, is clearly

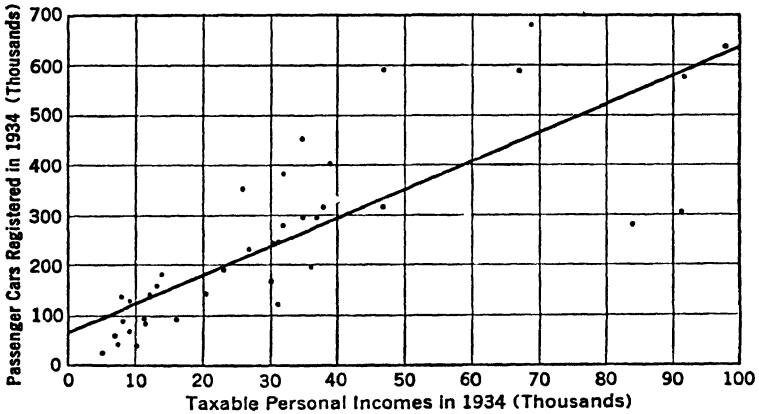


FIG. 69. — Scatter Diagram Showing the Relation between Taxable Personal Incomes and Passenger Car Registration, by States, in 1934, with Line of Average Relationship

existent. There is here a relationship analogous to a trend, and it is apparently a trend which can be represented by a straight line. The equation to a straight line, fitted by the method of least squares to the points on the scatter diagram, will express mathematically the *average relationship* between these two variables. Such a line could, of course, be fitted by inspection, but a more accurate result will be obtained by the method of least squares.

This calls for the solution of the following normal equations:

$$\Sigma(Y) = Na + b\Sigma(X)$$
$$\Sigma(XY) = a\Sigma(X) + b\Sigma(X^2).$$

The values required for the solution of these equations may be derived from the data as arranged in Table 86. Sub-

stituting, we have

$$9{,}549 = 38a + 1{,}242b$$
$$451{,}558 = 1{,}242a + 65{,}236b.$$

Solving

$$a = 66.321$$
$$b = 5.659.$$

The required equation is

$$Y = 66.321 + 5.659X.^{1}$$

This line is plotted in Fig. 69.

A mathematical expression has now been secured for the relation between the two variables being studied, the number of taxable personal incomes, by states, and the number of passenger automobiles registered. The former is the independent or X-variable in the equation, the latter the dependent or Y-variable. This equation constitutes a measure of the functional relationship between these two variables, but it is only an expression of *average* relationship. How significant is the equation? If the relationship were perfect, and the plotted points all lay on the line describing this relationship, the equation could be used with confidence as an accurate instrument for determining the value of one variable from a value of the other. But a line with a definite equation may be fitted to points which depart very widely from it, which are widely dispersed. In such a case the equation may have the appearance of describing a precise relationship but the variation is so great that it cannot be used with confidence. It is the same problem as that which arises when an average is employed. We must know how significant the average is, how great the concentration about it, before we may use it intelligently. So the equation of

[1] In the chapters on correlation capital letters (W, X, Y, etc.) are used to represent original values of the variable quantities, as measured from the zero points on the scales of actual values. Capital letters with prime marks are used to measure deviations from arbitrary origins, X' and Y' for such deviations in class-interval units, X'' and Y'' for such deviations in original units of measurement. Small letters (w, x, y, etc.) are used to represent values of the variables expressed as deviations from their respective arithmetic means.

relationship between variables means little unless we know to what extent it holds in practical experience. We must have a measure of the dispersion about the line we have fitted.

In describing the frequency distribution, the standard deviation is used as the best general measure of variation. It is, obviously, the measure we need in determining the reliability of the equation of average relationship. The standard deviation about this line will not only serve as a general index of the significance of this equation but will enable us to measure the degree of accuracy of estimates based upon the equation.

THE COMPUTATION OF THE STANDARD ERROR OF ESTIMATE

The standard deviation about a line of average relationship, being a measure of the accuracy of estimates, may be termed the *standard error of estimate.* The term *standard deviation* is generally confined to the root-mean-square deviation about the arithmetic mean. The standard error of estimate is represented by the symbol S.

In the computation of S we must know the computed value of Y which corresponds to each given value of X. By substituting the given values of X in the equation

$$Y = 66.321 + 5.659X$$

normal Y values may be computed. The deviations of the actual Y values from the computed may then be determined. The root-mean-square of these deviations is the required measure. A method of computation is illustrated in Table 87. From this table we have

$$S_y = \sqrt{\frac{421{,}398.31}{38}}$$

$$= 105.3 \text{ (thousand) motor cars.}$$

(The symbol S_y is used, as this is the standard error of the Y-variable.)

TABLE 87

Computation of Standard Error of Estimate

(1)	(2)	(3)	(4)	(5)
State	*No. of passenger cars registered,* 1934 *(in thousands)* *Y-actual*	*Y-computed*	*d* (2) − (3)	d^2
Alabama	192	196.5	− 4.5	20.25
Arizona	80	128.6	− 48.6	2,361.96
Arkansas	162	139.9	+ 22.1	488.41
Colorado	246	241.8	+ 4.2	17.64
Connecticut	310	581.3	− 271.3	73,603.69
Delaware	45	128.6	− 83.6	6,988.96
Florida	280	253.1	+ 26.9	723.61
Georgia	317	281.4	+ 35.6	1,267.36
Idaho	91	117.3	− 26 3	691.69
Indiana	680	462.4	+ 217 6	47,349.76
Iowa	591	337.9	+ 253.1	64,059.61
Kansas	453	270.0	+ 183.0	33,489.00
Kentucky	295	264.4	+ 30.6	936.36
Louisiana	199	275.7	− 76.7	5,882.89
Maine	141	185 2	− 44.2	1,953.64
Maryland	288	541.7	− 253.7	64,262.25
Minnesota	594	445.5	+ 148.5	22,052.25
Mississippi	141	139.9	+ 1.1	1.21
Missouri	632	620.9	+ 11.1	123.21
Montana	97	162.5	− 65.5	4,290.25
Nebraska	350	219.1	+ 130.9	17,134.81
Nevada	26	94.6	− 68.6	4,705.96
New Hampshire	91	162.5	− 71.5	5,112.25
New Mexico	67	111.6	− 44.6	1,989.16
N. Carolina	385	247.4	+ 137.6	18,933.76
N. Dakota	130	122.9	+ 7.1	50.41
Oklahoma	403	287.0	+ 116.0	13,456.00
Oregon	233	219.1	+ 13.9	193.21
Rhode Island	124	241.8	− 117.8	13,876.84
S. Carolina	182	151.2	+ 30.8	948.64
S. Dakota	146	111.6	+ 34.4	1,183.36
Tennessee	299	281 4	+ 17.6	309.76
Utah	85	128.6	− 43.6	1,900.96
Vermont	69	122.9	− 53 9	2,905.21
Virginia	317	338.0	− 21.0	441.00
W. Virginia	167	236.1	− 69.1	4,774.81
Wisconsin	589	592.6	− 3.6	12.96
Wyoming	52	105.9	− 53.9	2,905.21
Total				421,398.31

The measure S_y is to be interpreted in precisely the same way as the standard deviation about an arithmetic mean. Given an approximately normal distribution of items about the line of relationship, 68 per cent of all the cases will lie within a range of $\pm S$ (in this case 105.3), 95 per cent will fall within $\pm 2S$ (in this case 210.6) and 99.7 will fall within $\pm 3S$ (in this case 315.9). If there were no scatter about the line fitted to the points representing the corresponding values of X and Y, S would have a value of zero, and the value of Y could be estimated from the value of X with perfect accuracy. The less the dispersion about the line, the smaller the value of S. The value of S serves, therefore, as an indicator of the significance and usefulness of the line which describes the relationship between the two variables. The standard error, it should be noted, is expressed in the same units as the original Y-values.

THE MAKING OF ESTIMATES

We may, for a moment, consider the significance of these results. Let us assume that, not knowing the number of motor vehicles registered in a given state, we are under the necessity of estimating it. Two methods are open to us. We may, in the first place, base the estimate upon our knowledge of the Y-variable alone. The total number of passenger automobiles in the 38 states included in the study is 9,549,000. Dividing this by 38 we have 251,289 as the average. With no specific information as to the registration in a given state, the arithmetic mean of all the state figures would be taken as the most probable value for the state in question. (The most probable value of a series of observations is the mean of the series.) How may we judge of the accuracy of this estimate? The standard deviation of the original distribution is a measure of the degree of variation about the mean and, therefore, a measure of the accuracy of an estimate based upon the mean. If

the distribution approximates the normal type, the chances are 68 out of 100 that the true value for the state in question will not differ from the mean by more than the standard deviation. The standard deviation of passenger automobile registration by states, as recorded in Table 86, is 178.5. The mean affords, therefore, a basis for a reasonable estimate, and the standard deviation affords some indication of the probabilities involved in making this estimate.

Another method of estimating the motor vehicle registration in a given state is open to us if we know the number of taxable personal incomes in that state. We know, as a result of the study described in the preceding pages, that the average relationship between passenger car registration and number of taxable personal incomes is described by the equation

$$Y = 66.321 + 5.659X.$$

(The unit is 1,000 for each variable, it will be recalled.)

If a state has 50,000 taxable personal incomes, it may be estimated from this equation that there are 349,271 passenger automobiles registered in that state. This is the most probable value of Y as determined from the equation of average relationship. Is this estimate any better than the previous one, which took the mean Y as the most probable value? Does our knowledge of the average relationship between X and Y aid us in estimating the value of Y from a known value of X?

The answers to these questions are given by the *standard error of estimate*, and by the relationship between the standard error of estimate and the standard deviation of Y. The standard error of estimate (that is, the standard deviation about the line of average relationship) is 105.3. The standard deviation of Y is 178.5. Clearly the estimate made from the equation is more accurate than the estimate based upon the value of the mean Y. In the former case the odds are 68 out of 100 that the error will not exceed

105.3 or, in terms of the original units, 105,300 vehicles. When the estimate is made from the mean, the odds are 68 out of 100 that the error will not exceed 178,500 vehicles.[1] From our knowledge of the relationship between the two variables, even though that relationship is by no means constant or perfect, we are able to reduce materially the errors of estimate.

THE COEFFICIENT OF CORRELATION

We have now secured two measures which aid us in describing the relationship between variable quantities. The first is the fundamental equation of relationship, the expression of the degree of change in one variable associated, on the average, with a given change in the other. The second is the *standard error of estimate*, the measure of the degree of "scatter" about the line of average relationship. The standard error resembles the standard deviation in that it is a measure expressed in absolute terms, in the units employed in measuring the original Y-values. This measure enables us to determine in a given case the probability that an estimate based upon the equation of relationship will fall within certain limits.

In measuring variation it has been found that an abstract measure of variability is needed, one which is divorced from the absolute terms of the given problem. Such a measure is particularly needed, it was noted, when different distributions are to be compared. So, for measuring the *degree of variability*, a coefficient of variation is employed. There is need of a somewhat similar measure in connection with our present problem. We need a measure of the *degree of relationship* between two variables, an abstract coefficient which is divorced from the particular units

[1] In the present case, with a limited number of items and distributions which depart somewhat from the normal type, the precise probabilities cannot be so accurately determined from the values of S_y and σ_y. With this qualification in the matter of interpretation we may use S_y and σ_y as useful measures of dispersion.

employed in a given case. Karl Pearson has developed such a coefficient.

This measure may be explained in terms of the preceding discussion. It was found that the usefulness of estimates based upon the equation of relationship could be determined by comparing the standard error of Y (the measure of scatter about the line of relationship) with the standard deviation of Y. If the standard error be as great as the standard deviation the equation of relationship is of no use to us, but if the standard error be less than the standard deviation the accuracy of estimates may be improved by using this equation. The significance of the equation is thus indicated by the relation between the standard error and the standard deviation. But these are both in absolute terms, so that by dividing one by the other an abstract measure may be secured. Thus we might write

$$\text{Measure of correlation} = \frac{S_y}{\sigma_y}.$$

A somewhat more useful measure is secured by putting the ratio in this form:

$$\text{Measure of correlation} = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}.$$

This measure, when used in connection with a linear equation, is called the *coefficient of correlation* and is represented by the symbol r.

A brief consideration of this formula will help to make clear the significance of r. If there is no dispersion about the line of relationship, S_y will have a value of zero; the equation describes a perfect relationship between the two variables. In this case, as is clear from the formula, r must have a value of 1.

The maximum value of S_y is one which is equal to σ_y. Under these conditions, when the equation of relationship is of no aid in improving our estimates, the formula will

give zero as the value of r. Such a value indicates that there is no relationship between the two variables; in other words, that the straight line of best fit is horizontal, passing through the mean of the Y's. It shows that there is no tendency for the high values of Y to be associated with high values of X or for high values of Y to be associated with low values of X. The two variables fluctuate in absolute independence. In such a case the deviation of each point from the fitted line is equal to its deviation from the mean, and the two root-mean-square deviations are equal, as stated.

Zero and unity are thus the limits to the value of r. The values found in practical work fall somewhere between these limits, approaching unity in cases where the degree of relationship is high. The greater the value of r, the greater the confidence that may be placed in the equation as an expression of a relation which is approximated in a high percentage of cases. In the example presented above, dealing with motor vehicle registration and number of taxable personal incomes, we have

$$r = \sqrt{1 - \frac{(105.3)^2}{(178.5)^2}}$$
$$= .81.$$

This value indicates a definite and fairly close connection between these two variables for the states included in the sample.

The coefficient of correlation may be made somewhat more significant by giving it the sign of the constant b in the equation of relationship. This sign indicates whether the slope of the line is positive or negative and, when attached to r, enables us to tell whether the relationship is direct or inverse. Thus in the present case high values of one variable are paired with high values of the other. The correlation is positive and the coefficient should be written + .81. When cotton production and prices are

correlated the relationship is an inverse one, for high values of one variable are generally associated with low values of the other.

The measurement of relationship in a given case is completed when we have secured the three measures described. The *equation of average relationship* is an expression of the underlying law connecting the two variables, if such a law may be assumed. The *standard error of estimate* measures the variation, in absolute terms, about the line of relationship. The *coefficient of correlation* is an abstract measure of the degree to which the average relationship actually holds in practice.

DETAILS OF CALCULATION

In the preceding section the attempt has been made to explain the various measures necessary in studying the relationship between variable quantities without introducing a detailed explanation of procedure. We may now return to a consideration of the details of calculation, including certain methods by which this calculation may be reduced to a minimum.

The procedure followed in the preceding illustration is a logical one to employ in deriving the three required values. This method is capable of general application, but the labor involved may be materially reduced by taking advantage of a short-cut method of deriving S_y. This method may be first explained with reference to data of the type dealt with above. And, for the present, the discussion will be confined to cases in which the relationship between variables may be described by a straight line.

The first problem is the derivation of the equation of relationship. A line of the type

$$Y = a + bX$$

is fitted by the method of least squares.

The next step is the computation of S_y^2, the square of the

standard error of estimate. This was done in the above illustration by measuring the deviation of each individual observation from the fitted line, and getting the mean-square of these deviations. It may be shown[1] that this value can be derived from the following equation:

$$S_y^2 = \frac{\Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY)}{N}.$$

The quantities a and b are the constants in the equation to the fitted straight line. The other values relate to the original observations. Substituting in this equation a and b and the other necessary values, taken from Table 86, we have[2]

[1] The standard error of estimate is computed from the formula

$$S_y^2 = \frac{\Sigma(d^2)}{N}$$

where d represents a single deviation from the fitted line, or the difference between the actual and the computed value of Y in a given case. The latter is derived from the equation

$$Y_c = a + bX.$$

(The symbol Y_c is used to represent the computed value of Y.)

If we let Y represent the actual value, we have, for each residual,

$$d = Y_c - Y$$

or

$$d = a + bX - Y. \tag{1}$$

There will be as many equations of this type as there are points. Multiplying each one by d, and adding, we have

$$\Sigma(d^2) = a\Sigma(d) + b\Sigma(dX) - \Sigma(dY). \tag{2}$$

But, since the line was fitted by the method of least squares,

$$\Sigma(d) = 0$$
$$\Sigma(dX) = 0$$

(for proof of this see Appendix A)
and, therefore,

$$\Sigma(d^2) = -\Sigma(dY). \tag{3}$$

Returning again to equation (1), we may multiply throughout by Y, and add, securing

$$\Sigma(dY) = a\Sigma(Y) + b\Sigma(XY) - \Sigma(Y^2). \tag{4}$$

Substituting the equivalent of $\Sigma(dY)$ in equation (3), we have

$$\Sigma(d^2) = \Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY) \tag{5}$$

from which the given formula for S_y^2 is derived.

[2] For this calculation the values of a and b are given to a greater number of decimal places than in the equation as first presented.

$$S_y^2 = \frac{3{,}610{,}185 - (66.32136 \times 9{,}549) - (5.65925 \times 451{,}558)}{38}$$

$$= 11{,}090$$

$$S_y = 105.3.$$

From this point the procedure may follow that already described, r being computed from the formula

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}.$$

The coefficient r may be secured, however, without computing S as an intermediate value. The above formula for r may be reduced to

$$r^2 = \frac{a\Sigma(Y) + b\Sigma(XY) - Nc_y^2}{\Sigma(Y^2) - Nc_y^2}$$

where c_y is the difference between the mean Y and the origin employed in the calculations.[1] If the origin is zero

[1] The formula

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$

may be written

$$r^2 = 1 - \frac{\Sigma(d^2)}{\Sigma(y^2)}$$

in which y refers to deviations from the arithmetic mean of the Y's. But

$$\frac{\Sigma(y^2)}{N} = \frac{(\Sigma Y^2)}{N} - c_y^2$$

where Y represents a deviation from an arbitrary origin (in this case zero on the original scale) and c_y represents the difference between this origin and the mean of the Y's.

Therefore

$$r^2 = 1 - \frac{\Sigma(d^2)}{\Sigma(Y^2) - Nc_y^2}.$$

Substituting in this equation the equivalent of $\Sigma(d^2)$, as given in the footnote on page 338,

$$r^2 = 1 - \frac{\Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY)}{\Sigma(Y^2) - Nc_y^2}.$$

Simplifying,

$$r^2 = \frac{a\Sigma(Y) + b\Sigma(XY) - Nc_y^2}{\Sigma(Y^2) - Nc_y^2}.$$

on the original Y scale, c_y will be equal to the arithmetic mean of the Y's.

In the present case, using the data of Table 86, we have

$$c_y = \frac{9{,}549}{38} = 251.289.$$

The other values are the same as those employed above in computing S_y. Substituting in the formula, we have

$$r^2 = \frac{789{,}228.14}{1{,}210{,}630.86}$$
$$= .6519$$
$$r = .81.$$

In effect, then, the labor of fitting a straight line by the method of least squares gives us practically all the values needed in securing S and r, the two other measures necessary for a complete description of the relationship between two variable quantities. The only additional values required are $\Sigma(Y^2)$ and c_y.

The Construction of a Correlation Table

In the example presented above we had only thirty-eight observations. With a larger number it becomes practically impossible to retain the individual values in the study of relationships. These individual items must be grouped in significant classes, and all computations must be based upon these grouped data. This means, merely, that we must handle data organized in frequency distributions. Since we are dealing with two variables, however, the simple frequency table must be modified to meet the needs of the present problem. Such a modified frequency table, arranged to facilitate the computation of the values needed in studying relationship, is termed a *correlation table.*

As a typical problem, involving the construction of such a table, we may consider the relation between the discount rates of Federal Reserve banks and the corresponding dis-

count rates of commercial banks. Since the paper discounted by commercial banks may be rediscounted at the Federal Reserve banks by the member banks, some degree of relationship between the rates may be expected. Our present object is the measurement of that relationship.

The first step is the tabulation of the original observations. Monthly values of each variable[1] were secured for each of the twelve Federal Reserve cities over a period of 150 months, from July, 1920, to December, 1932. In the process of tabulation the items must be combined so that a Federal Reserve bank discount rate is paired with the corresponding rate charged by the commercial banks of the same city. Fig. 70 illustrates the method of tabulation.

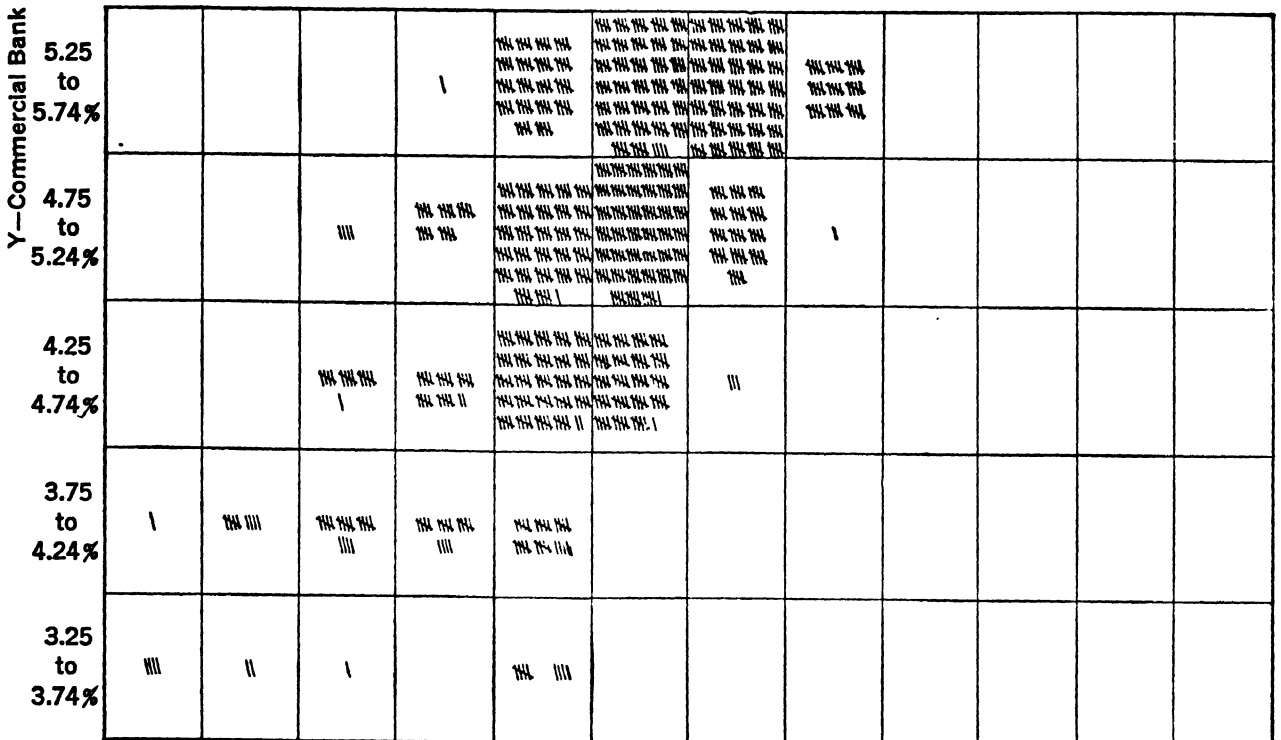
Tabulation having been completed, a correlation table designed to facilitate the later computations may be constructed. Table 88 illustrates a suitable form.

In Table 88, it will be noted, an arbitrary origin is employed for each variable, and the class-interval unit is used in the calculations. We here employ the symbols X' and Y' to represent deviations from the arbitrary origin (which is located at point 1.50, 3.50 on the original scales).

COMPUTATION OF MEASURES OF RELATIONSHIP

From this correlation table all the values needed in fitting a straight line to the data, and in computing the measures S and r, may be derived. The quantities $\Sigma(X')$, $\Sigma(X'^2)$, $\Sigma(Y')$, and $\Sigma(Y'^2)$ are computed by methods already familiar to the student. The product of the paired values $\Sigma(X'Y')$ may be computed directly from the correlation table, but it is perhaps simpler for the beginner to re-arrange the data in columnar form, as in Table 89 on page 345. When the figures are disposed in this way one line is em-

[1] The discount rates of the Federal Reserve banks relate, for the first part of the period covered, to trade acceptances; for later years they are "rates for member banks on eligible paper." The commercial bank rates are those charged on customers' prime commercial paper. The customary rate over a given 30 day period was taken as of the middle of that period.



342

FIG. 70. — Tabulation of Items in a Correlation Table

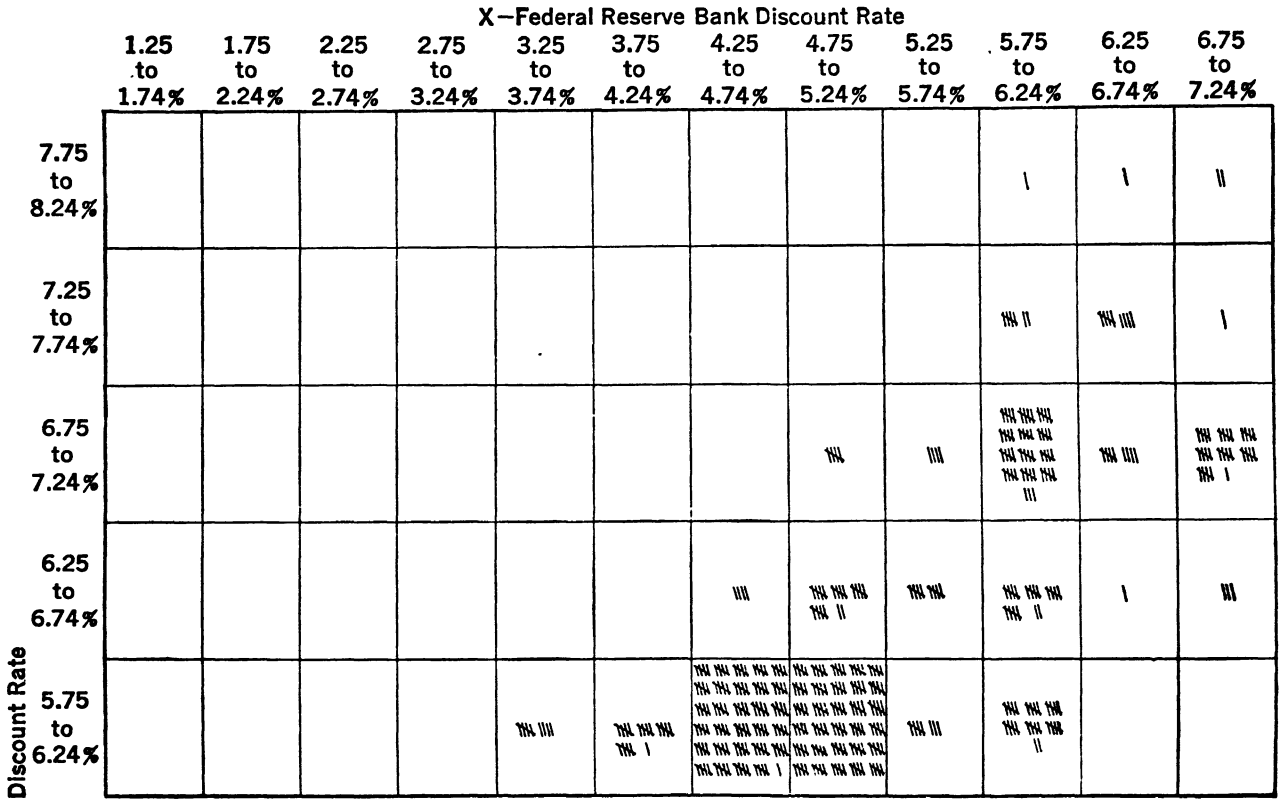

X—Federal Reserve Bank Discount Rate
1.25 to 1.74%
1.75 to 2.24%
2.25 to 2.74%
2.75 to 3.24%
3.25 to 3.74%
3.75 to 4.24%
4.25 to 4.74%
4.75 to 5.24%
5.25 to 5.74%
5.75 to 6.24%
6.25 to 6.74%
6.75 to 7.24%
7.75 to 8.24%
7.25 to 7.74%
6.75 to 7.24%
6.25 to 6.74%
5.75 to 6.24%
Discount Rate

TABLE 88

Correlation Table, Discount Rates of Federal Reserve Banks and Discount Rates of Commercial Banks

							X — Federal reserve banks' discount rates (per cent)												
Y — Discount rates of commercial banks (per cent)	*Class interval*						1.25–1.74	1.75–2.24	2.25–2.74	2.75–3.24	3.25–3.74	3.75–4.24	4.25–4.74	4.75–5.24	5.25–5.74	5.75–6.24	6.25–6.74	6.75–7.24	*Total*
		Mid-point					1.50	2.00	2.50	3.00	3.50	4.00	4.50	5.00	5.50	6.00	6.50	7.00	
			f				5	11	40	72	370	477	393	223	22	125	20	42	1,800
				d'			0	1	2	3	4	5	6	7	8	9	10	11	
					fd'		0	11	80	216	1,480	2,385	2,358	1,561	176	1,125	200	462	10,054
						fd'²	0	11	160	648	5,920	11,925	14,148	10,927	1,408	10,125	2,000	5,082	62,354
	7.75–8.24	8.00	4	9	36	324										1	1	2	
	7.25–7.74	7.50	17	8	136	1,088										7	9	1	
	6.75–7.24	7.00	117	7	819	5,733								5	4	63	9	36	
	6.25–6.74	6.50	62	6	372	2,232							4	22	10	22	1	3	
	5.75–6.24	6.00	366	5	1,830	9,150					9	21	146	150	8	32			
	5.25–5.74	5.50	475	4	1,900	7,600				1	90	164	175	45					
	4.75–5.24	5.00	402	3	1,206	3,618			4	25	111	196	65	1					
	4.25–4.74	4.50	264	2	528	1,056			16	27	122	96	3						
	3.75–4.24	4.00	77	1	77	77	1	9	19	19	29								
	3.25–3.74	3.50	16	0	0		4	2	1		9								
	Total		1,800		6,904	30,878													

TABLE 89

Discount Rates of Federal Reserve Banks and Discount Rates of Commercial Banks

(Computation of values required in curve fitting)

X'	Y'	f	$fX'Y'$
0	0	4	0
1	0	2	0
2	0	1	0
4	0	9	0
0	1	1	0
1	1	9	9
2	1	19	38
3	1	19	57
4	1	29	116
2	2	16	64
3	2	27	162
4	2	122	976
5	2	96	960
6	2	3	36
2	3	4	24
3	3	25	225
4	3	111	1,332
5	3	196	2,940
6	3	65	1,170
7	3	1	21
3	4	1	12
4	4	90	1,440
5	4	164	3,280
6	4	175	4,200
7	4	45	1,260
4	5	9	180
5	5	21	525
6	5	146	4,380
7	5	150	5,250
8	5	8	320
9	5	32	1,440
6	6	4	144
7	6	22	924
8	6	10	480
9	6	22	1,188
10	6	1	60
11	6	3	198
7	7	5	245
8	7	4	224

TABLE 89—*Continued*

Discount Rates of Federal Reserve Banks a d Discount Rates of Commercial Banks

X'	Y'	f	$fX'Y'$
9	7	63	3,969
10	7	9	630
11	7	36	2,772
9	8	7	504
10	8	9	720
11	8	1	88
9	9	1	81
10	9	1	90
11	9	2	198
			42,932

ployed for each compartment of the original correlation table in which items have been recorded.

The values required in fitting a straight line and in computing the standard error and the coefficient of correlation are:

$$N = 1{,}800 \qquad \Sigma(X'^2) = 62{,}354$$
$$\Sigma(X') = 10{,}054 \qquad \Sigma(fX'Y') = 42{,}932$$
$$\Sigma(Y') = 6{,}904 \qquad \Sigma(Y'^2) = 30{,}878.$$

The equation to the best fitting straight line is found to be

$$Y' = - .10277 + .70509X'.$$

Substituting in the formula

$$S_y^2 = \frac{\Sigma(Y'^2) - a\Sigma(Y') - b\Sigma(X'Y')}{N}$$

we have

$$S_y^2 = \frac{30{,}878 - (- .10277 \times 6{,}904) - (.70509 \times 42{,}932)}{1{,}800}$$
$$= .7314$$
$$S_y = .855.$$

To determine the value of the coefficient of correlation

we have only to substitute the proper values in the equation

$$r^2 = \frac{a\Sigma(Y') + b\Sigma(X'Y') - Nc_y^2}{\Sigma(Y'^2) - Nc_y^2}.$$

When this is done we have

$$r^2 = \frac{(-.10277 \times 6{,}904) + (.70509 \times 42{,}932) - (1{,}800 \times 14.71149)}{30{,}878 - (1{,}800 \times 14.71149)}$$

$$= \frac{3{,}080.7178}{4{,}397.3180}$$

$$= .70059$$

$$r = + .837.$$

All these calculations have been carried through in class-interval units, with reference to an origin at point 1.50, 3.50 on the original scales. The value of r is not affected by this fact, but the estimating equation and the standard error of estimate should be corrected.

The value of S_y, in class-interval units, is .855. Since the class-interval of the Y-variable is .50, we have, in original units,

$$S_y = .50 \times .855$$
$$= .4275.$$

The equation may be corrected in a similar fashion. The class-interval being .50 both for X and Y, each unit on the original scale equals two class-interval units. Thus a range of 4 points on the original scale is equivalent to a range of 8 points on the class-interval scale. For convenience we may use Y'' and X'' to define deviations in original units (i.e., deviations from the arbitrary origin), where we have used Y' and X' to define corresponding deviations in class-interval units. Then, for any stated deviation, $2Y'' = Y'$; similarly $2X'' = X'$. Retaining the values of a and b in the equation of average relationship, and substituting $2Y''$ for Y' and $2X''$ for X', we have

$$2Y'' = - .10277 + .70509(2X'').$$

Simplifying this, we have

$$Y'' = -.05138 + .70509X''$$

which is the equation in terms of original units.

This equation refers to an origin whose coördinates are 1.50 and 3.50 on the original scales. That is, $Y'' = Y - 3.50$, and $X'' = X - 1.50$, where Y and X define deviations, in original units, from the zero points on the original scales. Making these substitutions we have

$$Y - 3.50 = -.05138 + .70509(X - 1.50).$$

Simplifying, and rounding off the constants by dropping figures that are not significant, we have

$$Y = 2.391 + .705X.$$

We have now the three values required for determining the relationship between Federal Reserve discount rates

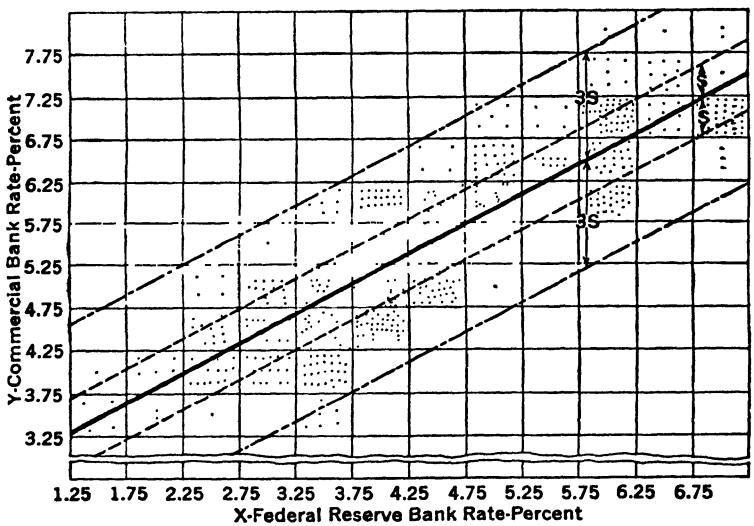


FIG. 71. — Scatter Diagram of Federal Reserve and Commercial Bank Rates, with Line of Average Relationship and Zones of Estimate

and corresponding commercial bank rates, during the period covered. The equation describes the average relationship, the standard error of estimate serves as a measure of the

reliability of estimates based upon this equation, and the coefficient of correlation serves as an abstract measure of the degree of relationship between the two variables.

The significance of the standard error, S_y, is brought out graphically in Fig. 71. The line of average relationship has been drawn on this scatter diagram, and what may be called "zones of estimate" have been marked out about this line. Within the zone having a width equal to $2S$, centering at the fitted straight line, 68 per cent of all the points should fall, on the assumption that the distribution is normal. Within the zone having a width equal to $6S$, centering at the fitted straight line, 99.7 per cent of all the points should fall, on the same assumption. The smaller the value of S the narrower these zones would be, and hence the more accurate would be the estimates which are based upon the equation of average relationship.

The Product-Moment Formula for the Coefficient of Correlation

In the preceding examples the coefficient of correlation has been computed from the formula

$$r^2 = \frac{a\Sigma(Y) + b\Sigma(XY) - Nc_y^2}{\Sigma(Y^2) - Nc_y^2}$$

which is based upon relations involved in fitting a straight line by least squares. The usual formula differs somewhat from this, and it is advisable that the student be familiar with it.

When a straight line is fitted to data, the origin being at the point of averages, the two normal equations

$$\Sigma(Y) = Na + b\Sigma(X)$$
$$\Sigma(XY) = a\Sigma(X) + b\Sigma(X^2)$$

become

$$\Sigma(y) = Na + b\Sigma(x)$$
$$\Sigma(xy) = a\Sigma(x) + b\Sigma(x^2)$$

where y and x measure deviations from the point of averages. The first of these equations disappears and the second reduces to

$$\Sigma(xy) = b\Sigma(x^2)$$

for

$$\Sigma(x) = 0 \text{ and } \Sigma(y) = 0.$$

The slope, b, is the only constant required, and this may be computed from the relationship

$$b = \frac{\Sigma(xy)}{\Sigma(x^2)}.$$

Under the same conditions the formula

$$r^2 = \frac{a\Sigma(Y) + b\Sigma(XY) - Nc_y^2}{\Sigma(Y^2) - Nc_y^2}$$

reduces to

$$r^2 = \frac{b\Sigma(xy)}{\Sigma(y^2)}$$

for $c_y = 0$ when the deviations are measured from the mean of the Y's. Substituting for b its equivalent, as just determined, we have

$$r^2 = \frac{\Sigma(xy) \cdot \Sigma(xy)}{\Sigma(y^2) \cdot \Sigma(x^2)}.$$

But $\Sigma(y^2) = N\sigma_y^2$ and $\Sigma(x^2) = N\sigma_x^2$.
Therefore

$$r^2 = \frac{\Sigma(xy) \cdot \Sigma(xy)}{N^2\sigma_y^2\sigma_x^2}$$

and

$$r = \frac{\Sigma(xy)}{N\sigma_x\sigma_y}$$

in which x and y refer to deviations from an origin at the point of averages

This formula may be expressed

$$r = \frac{p}{\sigma_x \sigma_y}$$

in which

$$p = \frac{\Sigma(xy)}{N}.$$

The quantity p is the *mean product* of the paired values of x and y.

The computation of the coefficient of correlation from this formula proceeds along lines somewhat different from those outlined above. As we have seen, both the arithmetic mean and the standard deviation may be readily computed by the selection of an arbitrary origin from which all deviations are measured, a later correction being made to offset the error involved in using this arbitrary origin. Similarly, the mean product p may be computed by a short method, requiring the use of assumed means and the application of a correction at the end of the process.

If x' and y' represent deviations from points arbitrarily selected as assumed means, while p' represents the mean product of such deviations, then

$$p' = \frac{\Sigma(x'y')}{N}.$$

The computation of p' is not difficult, for deviations may be measured from central points, and may be expressed in class-interval units. Having p' we may secure the true mean product from the formula

$$p = p' - c_x c_y$$

in which c_x and c_y represent the differences between the true and assumed means of the x's and y's, respectively.[1]

[1] The following is a proof of this relationship:

x' = deviation of any point from assumed mean of x's
x = deviation of same point from true mean of x's
c_x = difference between true and assumed means of x's
y' = deviation of same point from assumed mean of y's

(*Footnote 1 continued on page 352*)

The Product-Moment Method, Ungrouped Data

This method may be illustrated with reference, first, to ungrouped data, using the figures for personal incomes (X) and passenger car registration (Y), by states. The values required for this computation, as given in Table 86, are

$$\begin{aligned} N &= 38 \\ \Sigma(X) &= 1{,}242 \\ \Sigma(Y) &= 9{,}549 \\ \Sigma(X^2) &= 65{,}236 \\ \Sigma(Y^2) &= 3{,}610{,}185 \\ \Sigma(XY) &= 451{,}558. \end{aligned}$$

The mean product may be computed from the formula

$$p = \frac{\Sigma(xy)}{N} = \frac{\Sigma(x'y')}{N} - c_x c_y.$$

We may select as arbitrary origin the actual origin on the two original scales. Hence we have

$$p = \frac{\Sigma(XY)}{N} - c_x c_y.$$

(When the arbitrary origin is at zero on the original scales, the symbol X corresponds to x' and Y corresponds to y', as used in the formulas.)

For the two standard deviations

(*Footnote 1 continued from page 351*)

y = deviation of same point from true mean of y's
c_y = difference between true and assumed means of y's
$x' = x + c_x$
$y' = y + c_y$
$x'y' = (x + c_x)(y + c_y) = xy + c_x y + c_y x + c_x c_y.$

For the sum of all such products for N points, we have

$$\Sigma(x'y') = \Sigma(xy) + c_x\Sigma(y) + c_y\Sigma(x) + Nc_x c_y.$$

But $\Sigma(y) = 0$ and $\Sigma(x) = 0$.

Therefore

$$\Sigma(x'y') = \Sigma(xy) + Nc_x c_y$$

$$\frac{\Sigma(x'y')}{N} = \frac{\Sigma(xy)}{N} + c_x c_y$$

$$\frac{\Sigma(xy)}{N} = \frac{\Sigma(x'y')}{N} - c_x c_y$$

or $p = p' - c_x c_y.$

$$\sigma_x = \sqrt{\frac{\Sigma(X^2)}{N} - c_x^2}$$

$$\sigma_y = \sqrt{\frac{\Sigma(Y^2)}{N} - c_y^2}.$$

These measures may be computed readily from the values secured from Table 86:

$$c_x = \frac{1,242}{38} = 32.684 \qquad c_y = \frac{9,549}{38} = 251.289$$

$$c_x^2 = 1,068.2439 \qquad c_y^2 = 63,146.1615$$

$$p = \frac{451,558}{38} - 8,213.1297$$

$$= 3,670.0753$$

$$\sigma_x = \sqrt{\frac{65,236}{38} - 1,068.2439} \qquad \sigma_y = \sqrt{\frac{3,610,185}{38} - 63,146.1615}$$

$$= 25.47 \qquad = 178.49$$

Solving for the coefficient of correlation,

$$r = \frac{p}{\sigma_x \sigma_y} = \frac{3,670.0753}{25.47 \times 178.49} = + .8073.$$

The equation to the straight line which describes the average relationship between X and Y may be derived from the values required for the preceding calculations. When the origin is at the point of averages this equation may be written

$$y = r\frac{\sigma_y}{\sigma_x}x.$$

Substituting the proper values, we have

$$y = + .8073\,\frac{178.49}{25.47}\,x$$

$$= 5.657x.$$

This, with an insignificant difference, is the equation secured by the method of least squares. The constant term representing the y-intercept disappears, since the origin is at

the point of averages, through which the least squares line must pass.[1]

When the product-moment method is employed in computing the coefficient of correlation and in determining the equation of regression, the standard error, S_y, may be derived by a simple change in the formula first presented for r. From the expression

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$

we may secure the formula

$$S_y = \sigma_y\sqrt{1 - r^2}$$

which enables us to compute S_y, if we have the values of σ_y and r. In the present case,

$$\begin{aligned} S_y &= 178.49\sqrt{1 - .8073} \\ &= 105.3. \end{aligned}$$

The Product-Moment Method, Classified Data

The product-moment method is also applicable to cases in which it is necessary to construct a double frequency or

---

[1] That the formula $y = r\frac{\sigma_y}{\sigma_x}x$ is equivalent to the formula based upon the method of least squares may be readily demonstrated. When the line passes through the point of averages, the equation, $Y = a + bX$, becomes $y = bx$. But $b = \frac{\Sigma(xy)}{\Sigma(x^2)}$. We may write, accordingly, $y_c = \frac{\Sigma(xy)}{\Sigma(x^2)}x$.

This is equivalent to

$$y_c = r\frac{\sigma_y}{\sigma_x}x$$

for the latter may be written

(1) $y_c = \frac{\Sigma(xy)}{N\sigma_y\sigma_x} \cdot \frac{\sigma_y}{\sigma_x}x$ (3) $y_c = \frac{\Sigma(xy)}{N\sqrt{\frac{\Sigma(x^2)}{N}} \cdot \sqrt{\frac{\Sigma(x^2)}{N}}}x$

(2) $y_c = \frac{\Sigma(xy)}{N\sigma_x \cdot \sigma_x}x$ (4) $y_c = \frac{\Sigma(xy)}{\Sigma(x^2)}x$.

(The symbol y_c is employed for the computed value of y, in these equations, to avoid confusion with the actual y's which appear in the right-hand members of the equations.)

correlation table. The procedure is shown in detail in Table 90.

This table is identical with that previously presented for the same data, except that a different arbitrary origin has been selected.

The value 4.50 is adopted as the assumed mean of the X's (M'_x), and the value 5.50 as the assumed mean of the Y's (M'_y). Deviations are measured in class-interval units from this origin. In each compartment of the correlation table there are three figures, involved in the computation of $\Sigma(x'y')$. The figure in the center indicates the number of items falling in that compartment. Thus there are seven pairs having X values between 5.75 and 6.25 (mid-point 6.0) and Y values between 7.25 and 7.75 (mid-point 7.5). For each of these pairs x' (the deviation from the assumed mean of the X's) is + 3, in class-interval units, and y' (the deviation from the assumed mean of the Y's) is + 4, in class-interval units. For each pair, therefore, $x'y' = +12$. This figure appears at the top of the compartment. But there are seven pairs in this compartment, so the sum of $x'y'$ for this group is + 84. This figure appears in parentheses at the bottom of the compartment. To secure $\Sigma(x'y')$ for the entire table it is necessary to add algebraically the values secured in this way for all compartments. The addition is first carried out for the different rows, the sub-totals being given in the column at the right of the table. It is found that $\Sigma(x'y') = +4{,}492$, in class-interval units.

Details of the computation of the coefficient of correlation are given in Table 91 on page 358. The values of c_x and c_y are obtained by methods already familiar.

We have, from that table,

$$\frac{\Sigma(xy)}{N} = \frac{\Sigma(x'y')}{N} - c_x c_y$$
$$= +2.4277.$$

TABLE 90

Correlation Table Showing the Relation between Federal Reserve Bank Discount Rates and the Discount Rates of Commercial Banks, and Illustrating the Computation of Quantities Needed in the Measurement of Correlation

							X — *Discount rates of federal reserve banks (per cent)*												
Y — *Discount rates of commercial banks (per cent)*	*Class interval*						1.25–1.74	1.75–2.24	2.25–2.74	2.75–3.24	3.25–3.74	3.75–4.24	4.25–4.74	4.75–5.24	5.25–5.74	5.75–6.24	6.25–6.74	6.75–7.24	*Total*	$\Sigma(x'y)'$
		Mid-point					1.50	2.00	2 50	3.00	3.50	4.00	4.50	5.00	5.50	6.00	6.50	7.00		
			f				5	11	40	72	370	477	393	223	22	125	20	42	1,800	
				d'			−6	−5	−4	−3	−2	−1	0	+1	+2	+3	+4	+5		
					fd'		−30	−55	−160	−216	−740	−477	0	223	44	375	80	210	−746	
						fd'^2	+180	+275	640	648	1,480	+477	0	223	88	1,125	320	1,050	6,506	
	7.75–8.24	8.00	4	+5	+20	100										+15 1 (+15)	+20 1 (+20)	+25 2 (+50)		85
	7.25–7.74	7.50	17	+4	68	272										+12 7 (+84)	+16 9 (+144)	+20 1 (+20)		248
	6 75–7.24	7.00	117	+3	351	1,053								+3 5 (+15)	+6 4 (+24)	+9 63 (+567)	+12 9 (+108)	+15 36 (+540)		1,254

Y — Discount rates of																			
6.25–6.74	6.50	62	+2	124	248							0 4 (0)	+2 22 (+44)	+4 10 (+40)	+6 22 (+132)	+8 1 (+8)	+10 3 (+30)		254
5.75–6.24	6.00	366	+1	+366	366					−2 9 (−18)	−1 21 (−21)	0 146 (0)	+1 150 (+150)	+2 8 (+16)	+3 32 (+96)				223
5.25–5.74	5.50	475	0	0	0				0 1 (0)	0 90 (0)	0 164 (0)	0 175 (0)	0 45 (0)						
4.75–5.24	5.00	402	−1	−402	402			+4 4 (+16)	+3 25 (+75)	+2 111 (+222)	+1 196 (+196)	0 65 (0)	−1 1 (−1)						508
4.25–4.74	4.50	264	−2	−528	1,056			+8 16 (+128)	+6 27 (+162)	+4 122 (+488)	+2 96 (+192)	0 3 (0)							970
3.75–4.24	4.00	77	−3	−231	693	+18 1 (+18)	+15 9 (+135)	+12 19 (+228)	+9 19 (+171)	+6 29 (+174)									726
3.25–3.74	3.50	16	−4	−64	256	+24 4 (+96)	+20 2 (+40)	+16 1 (+16)		+8 9 (+72)									224
Total		1,800		−296	4,446														4,492

This is the value of p, the mean product, in class-interval units. Proceeding,

$$r = \frac{\Sigma(xy)}{N\sigma_x\sigma_y} = \frac{p}{\sigma_x\sigma_y}$$
$$= + .837.$$

In computing r, both the numerator and denominator of the final fraction (the mean product and the two standard deviations) are in class-interval units. Since this is true, r may be computed directly without reducing the figures to the original units. The entire operation, therefore, is carried on in simple class-interval units.

TABLE 91

Calculation of the Coefficient of Correlation between the Discount Rates of Commercial Banks and of Federal Reserve Banks

(Calculations based on the entries in Table 90)

$M'_x = 4.50$

$c_x = \frac{-746}{1{,}800} = - .414$

$c_x^2 = (- .414)^2 = .171$

$S_x^2 = \frac{6{,}506}{1{,}800} = 3.614$

$\sigma_x^2 = S_x^2 - c_x^2$
$= 3.614 - .171$
$= 3.443$

$\sigma_x = 1.855$

$M_x = 4.50 - .5(.414)$
$= 4.293$

$M'_y = 5.50$

$c_y = \frac{-296}{1{,}800} = - .164$

$c_y^2 = (- .164)^2 = .027$

$S_y^2 = \frac{4{,}446}{1{,}800} = 2.470$

$\sigma_y^2 = S_y^2 - c_y^2$
$= 2.470 - .027$
$= 2.443$

$\sigma_y = 1.563$

$M_y = 5.50 - .5(.164)$
$= 5.418$

$p = \frac{\Sigma(x'y')}{N} - c_x c_y$

$= \frac{4{,}492}{1{,}800} - [(- .414)(- .164)]$

$= 2.4956 - .0679$

$= + 2.4277$

$r = \frac{p}{\sigma_x\sigma_y}$

$= \frac{+ 2.4277}{(1.855)(1.563)}$

$= \frac{+ 2.4277}{2.8994}$

$r = + .837$

NOTE: The class-interval unit has been employed in all the computations shown in this table.

In deriving the equation to the straight line which describes the average relationship between x and y from the formula

$$y = r\frac{\sigma_y}{\sigma_x}x$$

σ_y and σ_x should be expressed in units of the original scales.[1]

[1] When the class-intervals happen to be the same, as in the present case, the change is not necessary, as the relation between numerator and denomi-

This is done by multiplying the present values by the class-intervals.

$$\sigma_x \text{ (in original units)} = 1.855 \times .50 = .9275$$
$$\sigma_y \text{ (in original units)} = 1.563 \times .50 = .7815.$$

Substituting the given values in the formula, we have

$$y = .837 \frac{.7815}{.9275} x$$
$$= .705x.$$

The Lines of Regression

In the above discussion certain terms ordinarily employed in the treatment of correlation have been purposely omitted. Several of these should be explained.

The equation to the line of best fit in the preceding illustration was found to be

$$y = .705x$$

when the origin was taken at the point of averages. In this equation y is expressed as a function of x; that is, x is taken to be the independent variable and y the dependent variable. The equation expresses the average variation in y (discount rates of commercial banks) corresponding to a change of one unit in x (discount rates of Federal Reserve Banks). This line of relationship corresponds precisely to a line of trend, which describes the average change in a given series accompanying a unit change in time. A line which thus describes the average relationship between two variables is termed a *line of regression.* Its equation is termed a *regression equation,* and the quantity $r \frac{\sigma_y}{\sigma_x}$ which gives the slope of such a line is called a *coefficient of regression.* The use of these terms dates back to early studies by Galton, dealing with the relation between the heights of

nator is not altered. In practice it is advisable always to express the two standard deviations in original units at this stage of the calculations.

fathers and the heights of sons. Sons, Galton found, deviated less on the average from the mean heights of the race than their fathers. Whether the fathers were above or below the average, the sons tended to go back or *regress* towards the mean. He therefore termed the line which graphically described the average relationship between these two variables the *line of regression.* The term is now used generally, as indicated above, though the original meaning has no significance in most of its applications.

In any given case equations to two lines of regression may be computed. One is an expression of the average relationship between a dependent Y-variable and an independent X-variable; the other describes the relationship between a dependent X-variable and an independent Y-variable. The significance of the two may be indicated graphically.

Figure 72 is derived directly from the scatter diagram presented in Fig. 71. The circle in each column represents the mean Y-value of all the items falling in that column. Thus in the third column there are 40 cases including all those with X-values falling between 2.25 per cent and 2.75 per cent. The Y-values vary, however, being distributed as shown in Table 92.

TABLE 92

Computation of the Arithmetic Mean of an Array

Class-interval	*Mid-point* m	*Frequency* f	fm
4.75 – 5.24	5.0	4	20.0
4.25 – 4.74	4.5	16	72.0
3.75 – 4.24	4.0	19	76.0
3.25 – 3.74	3.5	1	3.5
		40	171.5

$$M = \frac{171.5}{40} = 4.2875.$$

Similar mean values are obtained for the other columns.

These are plotted in Fig. 72, together with the line of regression of Y on X.

In Fig. 72 the X-variable (Federal Reserve bank discount rates) is independent. As it increases from 4.0 per cent to 4.5, 5.0, 5.5 per cent, and so on, the average of commercial bank rates increases also. An average commercial

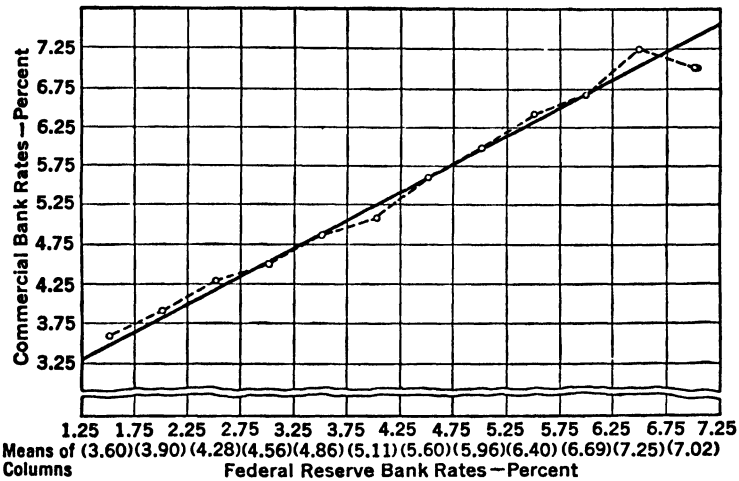


FIG. 72. — Showing the Relation between Discount Rates of Commercial Banks and Federal Reserve Bank Discount Rates. (The broken line connects the means of the columns and the straight line shows the average change in commercial bank rates corresponding to a unit change in Federal Reserve bank rates; i.e., it represents the regression of Y on X)

bank rate of 4.29 per cent was associated with an average Federal Reserve bank rate of 2.5 per cent; an average commercial bank rate of 4.56 per cent was associated with an average Federal Reserve bank rate of 3.0 per cent, and so on. (The commercial bank rates cited are the means of the entries in the columns referred to.) The slope of the straight line, which is the line of regression or the line of average relationship, measures the average increase in commercial bank rates corresponding to a unit increase in Federal Reserve bank rates.

It is possible to view the relationship between these two

variables in another light. These questions arise: Given a certain commercial bank discount rate, what is the average Federal Reserve bank rate associated with it? And for a given change in commercial bank discount rates, what is the average change in the corresponding Federal Reserve bank rates? The commercial bank rate is now looked upon as independent, and the Federal Reserve rate as an associated dependent variable. These questions are answered by Fig. 73. The points marked by the small circles and con-

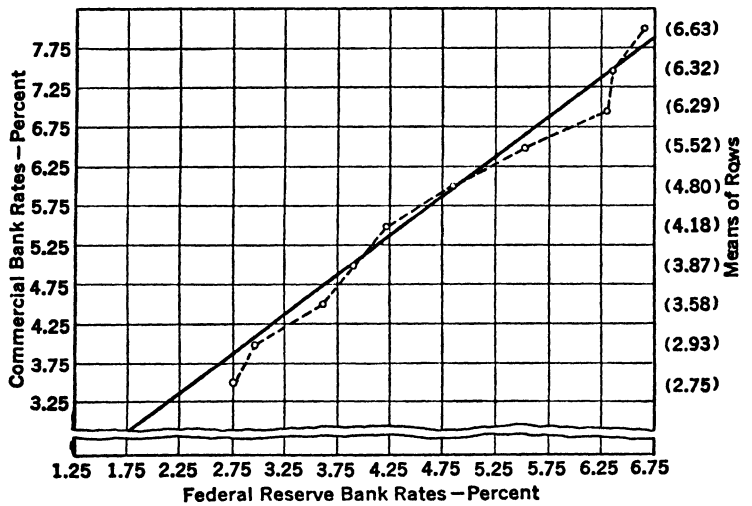


FIG. 73. — Showing the Relation between Federal Reserve Bank Discount Rates and the Discount Rates of Commercial Banks. (The broken line connects the means of the rows and the straight line shows the average change in Federal Reserve bank rates corresponding to a unit change in commercial bank rates; i.e., it represents the regression of X on Y)

nected by the broken line show the locations of the arithmetic means of the items falling in the various rows. Thus the 16 X-items in the bottom row have an average value of 2.75 per cent. This is the average Federal Reserve bank discount rate associated with a commercial bank rate of 3.5 per cent. The average Federal Reserve bank rate associated with a commercial bank rate of 4.0 per cent is 2.93 per cent, and so on. The straight line fitted to these

points indicates the relationship between the two, its slope measuring the average increase (or decrease) in Federal Reserve bank rates associated with a unit change in commercial bank rates.

This is the line of regression of X on Y. The general formula for the equation to this line is:

$$x = r\frac{\sigma_x}{\sigma_y}y.$$

Substituting the present values, we have

$$x = .837\,\frac{.9275}{.7815}\,y$$

or

$$x = .993y.$$

The factors in this equation, it will be seen, are the same as those entering into the formula for the line of regression of y on x[1]. If r is equal to 1 the two lines coincide, and if, in addition, the two standard deviations are equal, the line of regression will bisect the angle formed by the axes. If the points be plotted on a chart scaled in units of the standard deviations, we have $y = rx$; the slope of the line of regression is then equal to the value of r.

The coefficient of regression is represented by the symbol b. In a simple correlation problem there are two such coefficients, representing the slopes of the two lines of

[1] The formula $$x = r\frac{\sigma_x}{\sigma_y}y$$

may be reduced to $$x = \frac{\Sigma(xy)}{\Sigma(y^2)}y.$$

This is the equation to a line fitted to the points plotted in Fig. 73 in such a way that the sum of the squares of the *horizontal deviations* is a minimum.

The formula

$$y = \frac{\Sigma(xy)}{\Sigma(x^2)}x$$

is the equation to the line for which the sum of the squares of the *vertical deviations* is a minimum. An understanding of this point may make clear the difference between the two lines of regression.

regression. These are

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}.$$

(The subscripts indicate the relation between the two variables. The first subscript refers to the dependent variable in each case.)

The coefficient r appears in both formulas. This being so, it is clear that r may be computed from the regression coefficients. For

$$\sqrt{b_{yx} \cdot b_{xy}} = \sqrt{r\frac{\sigma_y}{\sigma_x} \cdot r\frac{\sigma_x}{\sigma_y}} = \sqrt{r^2} = r.$$

Thus if we know the slopes of the two lines of regression r may be determined. In the present example

$$r = \sqrt{.705 \times .993} = .837$$

USE OF THE EQUATIONS OF REGRESSION

The two equations of regression given above

$$y = .705x$$

and

$$x = .993y$$

describe relations between deviations from the respective arithmetic means. That is, the origin is at the point of averages, and to use the equations we cannot use the original values of X and Y but must express them as deviations from their means. For example, we wish to determine the normal commercial bank rate associated with a Federal Reserve bank rate of 6 per cent. The mean value of the X-variable (Federal Reserve bank rates) is 4.293 per cent. A rate of 6 per cent represents a deviation from the mean of + 1.707. Substituting this value in the first of the above equations, we have

$$y = .705 \times (+ 1.707)$$
$$= + 1.203.$$

This is the average y-deviation associated with an x-deviation of $+ 1.707$. To get the normal commercial bank rate associated with a Federal Reserve rate of 6 per cent the quantity $+ 1.203$ per cent must be added to the mean commercial bank rate, 5.418 per cent. The value we wish is thus 6.621 per cent.

This calculation has been rather round-about because of the form of the equation of relationship. This equation can be put in more appropriate form for such computations

Let

$\overline{X}$ = arithmetic mean of the X's
$\overline{Y}$ = arithmetic mean of the Y's.

Then

$$y = r\frac{\sigma_y}{\sigma_x}x$$

may be written

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X}).$$

In this last equation X and Y represent the values of the variables on the original scales, and not as deviations from their respective means. In terms of the coördinate chart, it means shifting the origin from the point of averages to a point corresponding to zero on each of the original scales.

To illustrate the greater utility of the equation in this form, the equation

$$y = .705x$$

may be changed in the manner indicated. It becomes

$$Y - 5.418 = .705(X - 4.293)$$
$$= .705X - 3.027$$
$$Y = 2.391 + .705X.$$

This is the equation with the origin so shifted that the

original values may be employed directly. To determine the commercial bank rate normally associated with a Federal Reserve rate of 6 per cent we may substitute the latter value in the equation just secured.

$$\begin{aligned} Y &= 2.391 + (.705 \times 6.0) \\ &= 6.621. \end{aligned}$$

Precisely the same results are secured as with the equation in the other form, but for many purposes it is preferable to have an equation in which the actual values may be inserted.

The equation

$$x = r\frac{\sigma_x}{\sigma_y}y$$

may be similarly changed to

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y}).$$

Summary of Correlation Procedure

In the foregoing pages there have been presented two quite different methods of securing the values required in measuring the relationship between two variables. The steps in the two methods may be briefly summarized. The method of least squares is basic in both cases, but that term may appropriately be employed to describe the first method outlined, for the process of fitting the line is the first and fundamental step in that procedure.

I. The Least Squares Method.

A. Data to be handled as individual items.

1. Fit a straight line to the data by the method of least squares. A simple arrangement of the data in columns will permit the ready computation of the required values, $\Sigma(X)$, $\Sigma(Y)$, $\Sigma(X^2)$, $\Sigma(Y^2)$, $\Sigma(XY)$. The equation thus obtained describes the average relationship between the two variables.

2. Compute the standard error of estimate, S_y, from the formula

$$S_y^2 = \frac{\Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY)}{N}.$$

S_y is a measure of the reliability of estimates based upon the equation of relationship, and is to be interpreted in the same way as is the standard deviation about an arithmetic mean.

3. Compute the coefficient of correlation, r, from the formula

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$

or from

$$r^2 = \frac{a\Sigma(Y) + b\Sigma(XY) - Nc_y^2}{\Sigma(Y^2) - Nc_y^2}.$$

Give r the sign of the constant b in the equation of regression. This coefficient is an abstract measure of the degree of relationship between the two variables, in so far as this relationship may be described by a straight line.

4. If an equation describing the regression of X on Y (X being dependent) is desired, the proper values may be substituted in the two normal equations

$$\Sigma(X) = Na + b\Sigma(Y)$$
$$\Sigma(XY) = a\Sigma(Y) + b\Sigma(Y^2).$$

The equation secured will be of the type

$$X = a + bY.$$

The standard error of estimate, S_x, may be computed by making the appropriate changes in the formula as given for S_y. The value of r will be the same as in the preceding case, in which Y is dependent.

B. Data to be classified.

1. Select an appropriate class-interval and tabulate the items in the form of a correlation table.
2. Compute the necessary values for fitting a straight line to the data. In doing so, an arbitrary origin may be selected for each variable, and all values expressed in class-interval units. A re-arrangement in columnar form may facilitate the computation of the quantity

$$\Sigma(X'Y').$$

3. Compute the standard error of estimate, employing the formula given above.
4. Compute the coefficient of correlation from the formula given above.
5. If the above calculations were carried on in class-interval units, the equation of average relationship and the standard error of estimate should now be expressed in terms of the original units of measurement. If an arbitrary origin was employed, the equation should be corrected so that the variables relate to deviations from the true origin.

II. The Product-Moment Method.

A. Data to be handled as individual items.

1. Arrange the paired observations in parallel columns and compute the quantities $\Sigma(X)$, $\Sigma(Y)$, $\Sigma(X^2)$, $\Sigma(Y^2)$, $\Sigma(XY)$.
2. Divide these quantities throughout by N. For the first two of these quotients we may use the symbols c_x and c_y (i.e.,

$$\frac{\Sigma(X)}{N} = c_x$$

and

$$\frac{\Sigma(Y)}{N} = c_y).$$

3. Compute the mean product from the formula

$$p = \frac{\Sigma(XY)}{N} - c_x c_y.$$

4. Compute the two standard deviations from the formulas

$$\sigma_x = \sqrt{\frac{\Sigma(X^2)}{N} - c_x^2}$$

$$\sigma_y = \sqrt{\frac{\Sigma(Y^2)}{N} - c_y^2}.$$

5. Compute the coefficient of correlation from the formula

$$r = \frac{p}{\sigma_x \sigma_y}.$$

6. Determine the equations of regression by substituting the proper values in the formulas

$$y = r\frac{\sigma_y}{\sigma_x}x$$

$$x = r\frac{\sigma_x}{\sigma_y}y.$$

(Note: For each of these equations the origin is at the point of averages.)

7. If desired, transfer the origin to zero on the two original scales by substituting the arithmetic means in the equations

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$$

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y}).$$

8. Compute the two standard errors of estimate from the formulas

$$S_y = \sigma_y\sqrt{1 - r^2}$$

$$S_x = \sigma_x\sqrt{1 - r^2}.$$

B. Data to be classified.

1. Construct a correlation table as in I. B. above.
2. Select an assumed mean for each variable. Measure the deviations of the various items from the assumed means in class-interval units.
3. Compute c_x and c_y in class-interval units.
4. Compute σ_x and σ_y in class-interval units.
5. Compute $\Sigma(x'y')$ in class-interval units for each compartment of the correlation table. Total these figures to get $\Sigma(x'y')$ for the whole table.
6. Determine the value of the mean product in class-interval units from the formula

$$p = \frac{\Sigma(x'y')}{N} - c_xc_y.$$

7. Compute r from the formula

$$r = \frac{p}{\sigma_x\sigma_y}.$$

8. Reduce σ_x and σ_y to original units.
9. Determine the equations of regression by substituting the proper values in the formulas

$$y = r\frac{\sigma_y}{\sigma_x}x$$

and

$$x = r\frac{\sigma_x}{\sigma_y}y.$$

10. If desired, transfer the origin to zero on the two original scales from the formulas

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$$

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y}).$$

11. Compute the two standard errors of estimate from the formulas

$$S_y = \sigma_y\sqrt{1 - r^2}$$

$$S_x = \sigma_x\sqrt{1 - r^2}.$$

It is advisable, in all cases, to construct scatter diagrams and to plot the lines of regression thereon. It is generally possible to derive from such diagrams a truer idea of the relations involved, and of the adequacy of the methods employed, than may be obtained from a study of the figures alone.

LIMITATIONS

A question naturally arises as to the degree of generality attaching to the measures of relationship described in the preceding pages. Are they limited to certain types of distributions, or may they be employed as absolutely general and universally valid measures?

As we have seen, the standard deviation has a precise and definite meaning with respect to distributions following the

normal law. Having values of the mean and of the standard deviation, we know, in such cases, the exact percentage of observations which will fall within any stated limits. If the distribution departs from the normal type the standard deviation is still a useful measure, but it cannot be interpreted in the same exact sense. Bearing this in mind, the formula

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$

may be considered.

When the distribution of the original values of the dependent variable about their mean is normal and the distribution about the least squares line is normal, both S_y and σ_y have specific and exact meanings, and it is perfectly legitimate to compute such a measure as r, based upon the relation of one to the other. Departures from normality in either case reduce the significance of this comparison. But we have seen that the standard deviation remains a useful measure even though the departure from the normal type be fairly pronounced, though in the latter case it lacks the precise significance attaching to it in a normal distribution. In the same way the standard error of estimate and the coefficient of correlation may be computed and utilized, even when all the requirements of normality are not met. Care must be taken in their interpretation in such cases, however. It must be clearly recognized that these measures have their full significance only in cases where the original distribution of the dependent variable and the distribution about the least squares line are both normal, or approximately so.

A simple example may make clear the effect upon the value of the coefficient of correlation of an extreme departure from a normal distribution. In the following table are listed certain selected figures taken from the 1919 Census of Manufactures, for the State of New York.

TABLE 93

Wage-Earners in Factories and Value of Products, 1919, in Eleven Cities in the State of New York

City	*Number of wage-earners (in thousands)* (*X*)	*Total value of products (in millions of dollars)* (*Y*)
Batavia	2.2	9
Beacon	2.2	10
Corning	3.5	11
Geneva	2.5	10
Glens Falls	2.8	12
Ithaca	1.7	10
Middletown	2.2	10
Peekskill	2.1	11
Rensselaer	1.4	10
Tonawanda	1.8	16
New York City	638.8	5,261

When the first ten of these cities, in the order listed, are treated as a group, the following values are secured:

$$\sigma_y = 1.8682$$
$$S_y = 1.8669$$
$$r = -.034.$$

(No general significance is to be attached to the above coefficient of correlation, for the cities were selected for the purpose of illustrating a particular point.)

The ten points and the line of regression are plotted in Fig. 74.

When New York City is included in the group, the values secured for the sample of eleven cities are

$$\sigma_y = 1509.3$$
$$S_y = 7.53$$
$$r = +.999988.$$

The eleven points and the line of regression are plotted in Fig. 75.

The reason for the markedly different results is obvious.

When the one very large city is included with the ten small cities the standard deviations of both variables are greatly increased. That of the Y-variable (value of products) is increased from 1.8682 to 1509.3. But S_y, the measure of the scatter about the fitted line, undergoes no such pro-

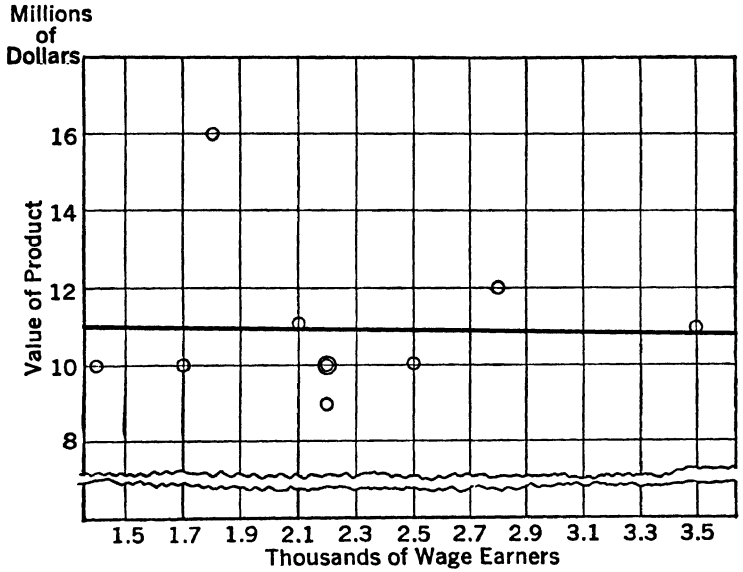


FIG. 74. — Showing the Relation between Number of Wage-Earners in Factories and Value of Products in Ten Selected Cities in the State of New York

nounced change in value. For the ten cities it is 1.8669; for the eleven cities 7.53. This is due to the fact that the one exceptional case is given such a great weight, in fitting by the method of least squares, that the fitted line must pass through or very near the point representing this observation. Accordingly, S is always affected less than σ by a single very exceptional case. Since the value of r depends upon the relationship

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}},$$

the presence of such a case always tends to increase the

value of the measure of correlation. The introduction of the one exceptional case in the above example changes a correlation coefficient of virtually zero to one of unity. The result, of course, is meaningless.

While this example represents an extreme instance, the same distortion will be felt, to a greater or less degree, whenever there is a departure from a normal distribution.

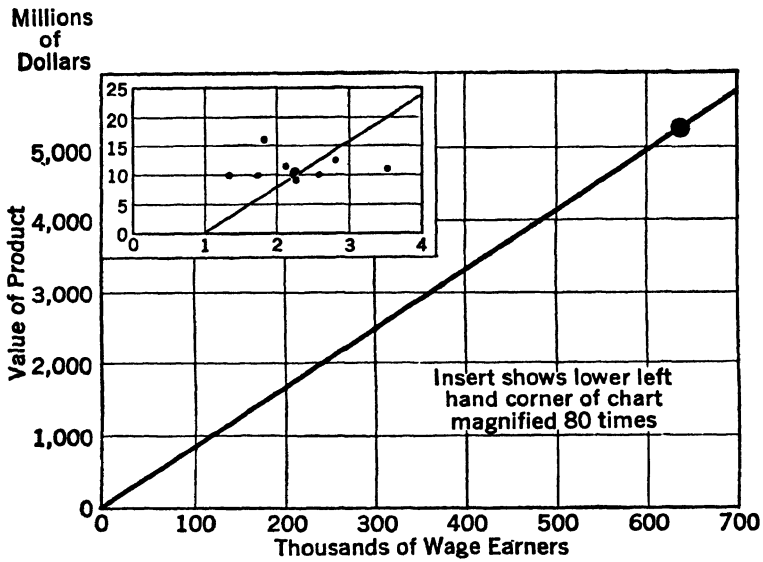


FIG. 75. — Showing the Relation between Number of Wage-Earners in Factories and Value of Products in Eleven Selected Cities in the State of New York

In practice the various measures of relationship cannot be restricted to perfectly normal distributions, but they must be interpreted with care when there is reason to believe that such disturbing influences are present.

THE COEFFICIENT OF RANK CORRELATION

The coefficient of rank correlation is a measure of relationship not subject to the distortion introduced by material departures from normality, and one which is particularly

useful in providing an objective test of the existence of correlation. Its application calls merely for the orderly ranking of observations. Thus we may rank 47 states of the union[1] according to the number of individual income tax returns in 1934, and according to the number of passenger automobiles registered in that year. The results are shown in Table 94.

TABLE 94

Illustrating the Computation of the Coefficient of Rank Correlation

(1)	(2)	(3)	(4)	(5)
State	*Rank on basis of number of individual income tax returns in* 1934	*Rank on basis of number of passenger automobiles registered in* 1934	*Difference* (2) − (3) *d*	d^2
Nevada	1	1	0	0
Wyoming	2	3	− 1	1
New Mexico	3	4	− 1	1
S. Dakota	4	15	− 11	121
Idaho	5	9	− 4	16
N. Dakota	6	12	− 6	36
Vermont	7	5	+ 2	4
Delaware	8	2	+ 6	36
Arizona	9	6	+ 3	9
Utah	10	7	+ 3	9
Mississippi	11	13	− 2	4
Arkansas	12	16	− 4	16

[1] Washington is excluded, because the published income tax returns for that state include those of Alaska.

Following are the records for the nine states not listed in Table 86.

	No. of taxable personal incomes (in thousands) 1934	*No. of passenger automobiles registered (in thousands)* 1934
California	316	1,769
Illinois	310	1,282
Massachusetts	243	687
Michigan	139	1,026
New Jersey	211	741
New York	808	1,971
Ohio	210	1,453
Pennsylvania	342	1,466
Texas	119	1,086

TABLE 94—*Continued*

Illustrating the Computation of the Coefficient of Rank Correlation

(1)	(2)	(3)	(4)	(5)
State	*Rank on basis of number of individual income tax returns in* 1934	*Rank on basis of number of passenger automobiles registered in* 1934	*Difference* (2) − (3) d	d^2
S. Carolina	13	18	− 5	25
New Hampshire	14	8	+ 6	36
Montana	15	10	+ 5	25
Maine	16	14	+ 2	4
Alabama	17	19	− 2	4
Nebraska	18	30	− 12	144
Oregon	19	21	− 2	4
W. Virginia	20	17	+ 3	9
Colorado	21	22	− 1	1
Rhode Island	22	11	+ 11	121
N. Carolina	23	31	− 8	64
Florida	24	23	+ 1	1
Kentucky	25	25	0	0
Kansas	26	33	− 7	49
Louisiana	27	20	+ 7	49
Tennessee	28	26	+ 2	4
Georgia	29	29	0	0
Oklahoma	30	32	− 2	4
Virginia	31	28	+ 3	9
Iowa	32	35	− 3	9
Minnesota	33	36	− 3	9
Indiana	34	38	− 4	16
Maryland	35	24	+ 11	121
Connecticut	36	27	+ 9	81
Wisconsin	37	34	+ 3	9
Missouri	38	37	+ 1	1
Texas	39	42	− 3	9
Michigan	40	41	− 1	1
Ohio	41	44	− 3	9
New Jersey	42	40	+ 2	4
Massachusetts	43	39	+ 4	16
Illinois	44	43	+ 1	1
California	45	46	− 1	1
Pennsylvania	46	45	+ 1	1
New York	47	47	0	0
				1,094

The degree of correlation is indicated by the degree of concordance between the two rankings. A precise measure of correlation is provided by the coefficient

$$\rho_r = 1 - \frac{6\Sigma d^2}{n^3 - n}$$

where d is a difference between the rankings of a given state in columns (2) and (3), and n is the number of states included.[1] (The Greek letter rho (ρ) with subscript r is used as the symbol of this coefficient.)

The method of computation is shown in Table 94. From the measurements there given we have

$$\rho_r = 1 - \frac{6 \times 1{,}094}{(47)^3 - 47} = 1 - \frac{6{,}564}{103{,}776}$$
$$= .94.$$

[1] This formula may be derived from the familiar product-moment formula for the coefficient of correlation, simplified because of the fact that the sums of the squares of the deviations of the first n natural numbers from their mean is equal to $\frac{n^3 - n}{12}$.

If we let d equal the difference between the rank of one variable and the corresponding rank of the other, we have, for any given pair of observations,

$$d = X - Y = x - y \quad \text{(since the means of the two series of ranks are identical)}$$
$$\Sigma d^2 = \Sigma(x - y)^2 = \Sigma x^2 + \Sigma y^2 - 2\Sigma xy$$
$$2\Sigma xy = \Sigma x^2 + \Sigma y^2 - \Sigma d^2.$$

But $\Sigma x^2 = \frac{n^3 - n}{12}$ and $\Sigma y^2 = \frac{n^3 - n}{12}$.

Therefore $2\Sigma xy = \frac{2n^3 - 2n}{12} - \Sigma d^2$

$$\Sigma xy = \frac{1}{2}\left(\frac{n^3 - n}{6} - \Sigma d^2\right).$$

But $\rho_r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$ (the product-moment formula for r)

$$= \frac{\frac{1}{2}\left(\frac{n^3 - n}{6} - \Sigma d^2\right)}{\frac{n^3 - n}{12}}$$
$$= 1 - \frac{6\Sigma d^2}{n^3 - n}.$$

The coefficient of rank correlation is appropriate where it is possible to rank individuals, or other entities, on the basis of abilities or qualities not open to exact measurement. It is also well adapted for use where the distributions of the observations depart widely from the normal type, and where the usefulness of customary measurements would be seriously impaired. This point takes on particular importance in connection with tests of significance, involving generalizations from sample results.[1] Such tests are discussed in later chapters.

REFERENCES

Bowley, A. L., *Elements of Statistics*, Part II, Chaps. 6, 7.

Brunt, David, *The Combination of Observations*, Chap. 10.

Burgess, Robert, *The Mathematics of Statistics*, Chap. 10.

Camp, B. H., *The Mathematical Part of Elementary Statistics*, Part I, Chaps. 8, 9.

Chaddock, R. E., *Principles and Methods of Statistics*, Chap. 12.

Croxton, F. E. and Cowden, D. J., *Practical Business Statistics*, Chap. 19.

Crum, W. L. and Patton, A. C., *An Introduction to the Methods of Economic Statistics*, Chaps. 15, 16.

Davies, G. R. and Yoder, Dale, *Business Statistics*, Chap. 6.

Day, E. E., *Statistical Analysis*, Chaps. 12, 13.

Elderton, W. P., *Frequency Curves and Correlation*, Chap. 7.

Ezekiel, Mordecai, *Methods of Correlation Analysis*, Chaps. 1–5, 7–9.

Florence, P. S., *The Statistical Method in Economics and Political Science*, Chap. 9.

Galton, Francis, "Correlations and their Measurement." *Proceedings of the Royal Society*, Vol. XLV, 1888 (136–145).

Jones, D. C., *A First Course in Statistics*, Chap. 10.

Kelley, Truman L., *Statistical Method*, Chap. 8.

Moore, H. L., *Forecasting the Field and the Price of Cotton*, Chap. 2.

Pearl, Raymond, *Introduction to Medical Biometry and Statistics*, Chap. 14.

[1] See Harold Hotelling and Margaret Pabst, "Rank Correlation and Tests of Significance Involving No Assumption of Normality." *Annals of Mathematical Statistics*, Vol. 7, 1936, 29–43.

Pearson, Karl, *The Grammar of Science*, Chaps. 4, 5.

Pearson, Karl, "Regression, Heredity and Panmixia." *Philosophical Transactions*, Royal Society. Series A. Vol. CLXXXVII, 1896 (253–318).

Richardson, C. H., *An Introduction to Statistical Analysis*, Chap. 7.

Rietz, H. L. and Crathorne, A. R., "Simple Correlation" (in *Handbook of Mathematical Statistics*, Rietz, H. L. ed., Chap. 8.)

Smith, J. G., *Elementary Statistics*, Chaps. 20, 21.

Snedecor, G. W., *Statistical Methods Applied to Experiments in Agriculture and Biology*, Chaps. 6, 7.

Tippett, L. H. C., *The Methods of Statistics*, Chap. 7.

Walker, Helen M., *Studies in the History of Statistical Method*, Chap. 5.

Waugh, A. E., *Elements of Statistical Method*, Chap. 9.

Whittaker, E. T. and Robinson, G., *The Calculus of Observations*, Chap. 12.

Yule, G. U. and Kendall, M. G., *An Introduction to the Theory of Statistics*, Chaps. 11, 12.

CHAPTER XI

THE MEASUREMENT OF RELATIONSHIP BETWEEN TIME SERIES

The methods of measuring correlation described in the preceding chapter were devised originally for the analysis of non-historical data, that is, for the treatment of *frequency* series rather than *time* series. The measurement of correlation between series in time presents certain distinctive problems which require separate treatment.

We have seen that such series are affected by various forces, which have been classified as the secular trend, cyclical and seasonal fluctuations and accidental variations, and methods have been described by means of which the effects of these various forces may be isolated. This breaking up of a series into its component parts for separate study is essential in attempting to correlate series in time, for spurious and quite misleading results will be secured if this is not done. The problem of correlation is that of securing a precise measure of the degree of relationship between variable quantities. But each series in time represents the combination of a number of variables and, so far as possible, each should be treated separately in correlating such series.

The relationship between two time series as, for example, interest rates and bond prices, may be studied with respect to any or all of the following components:

a. Secular trend.
b. Cyclical fluctuations.
c. Seasonal fluctuations.
d. Changes from one time unit to the next (e.g., week to week, month to month, or year to year).

Such relationships may be studied, first, through the comparison of graphs, and much may be learned by this simple process. The similarity or dissimilarity of secular trends, and the general relation between cyclical movements may be determined by a study of such graphs. For more accurate comparison the coefficient of correlation may be used, but when it is so employed it is particularly important that the precise nature of its employment and the exact significance of the results be understood.

For the comparison of secular trends the coefficient of correlation would never be employed. The mere fact that two series have the same secular trend is no indication of a relationship of interdependence; a coefficient of correlation based upon the trend values would be meaningless. Moreover, much simpler methods are available for comparing trends.

For the same reason a coefficient of correlation should not be based upon the original absolute values of two series in time, except in the rather rare case in which neither series is marked by a definite secular trend. The computation of r, when dealing with ordinary statistical data, involves measuring the deviations of all the items from their respective arithmetic means, and securing the sum of the products of the paired deviations. When deviations of like sign are paired throughout r will have a positive value; when deviations of unlike signs are paired throughout r will have a high negative value. The presence of pronounced rising or declining secular trends makes it impossible to secure significant values for r by the employment of this method. For example, the relation between automobile production and the price of bacon between the years 1900 and 1920 might be measured. The secular trend is markedly rising in each case. When the deviations of the annual figures are measured from the arithmetic means of the two series, the paired items for the earlier years will be negative, for the later years positive. A fairly high positive

value for r would be secured, were the computation carried through on this basis. This value would be quite misleading, for no real relationship can be expected in this case. The coefficient of correlation in such a case would measure, primarily, the relation between the two secular trends.

This coefficient might conceivably be employed to determine the similarity between seasonal fluctuations in two series, but its utility for this purpose may be questioned. Here again other and simpler methods are available.

In practice, therefore, the device of correlation should be employed neither to measure the relation between secular trends nor between seasonal movements. Its use is confined to comparisons of two or more series with respect to cyclical fluctuations and with respect to the short time changes from month to month or year to year. And, if valid measures of correlation are to be secured in making such comparisons, the effects of forces which distort these comparisons should be eliminated, in so far as this is possible. The actual work of correlation must be preceded by a sifting process designed to remove such irrelevant material. Unless the data are thus "distilled" the interpretation of the resulting coefficients will be difficult.

The Measurement of Correlation between Cyclical Fluctuations

In an earlier chapter we have dealt with methods by which the effects of certain of the factors affecting time series might be measured and eliminated. The spurious correlation due to secular trend may be avoided by measuring the deviations of the observations not from the respective arithmetic averages but from the lines of secular trend of the two series. These variations, the deviations from trend, are the significant values if our interest centers in the cycles. If annual values are employed the problem of eliminating seasonal fluctuations is not faced.

To illustrate this method of measuring the relationship

between series in time we may undertake to determine whether there is any connection between cyclical fluctuations in cotton production and in cotton prices. Figures for crop years are to be employed, for the period 1901–02 to 1935–36.

Cotton prices require some correction before correlation is attempted. The raw figures with which the investigation

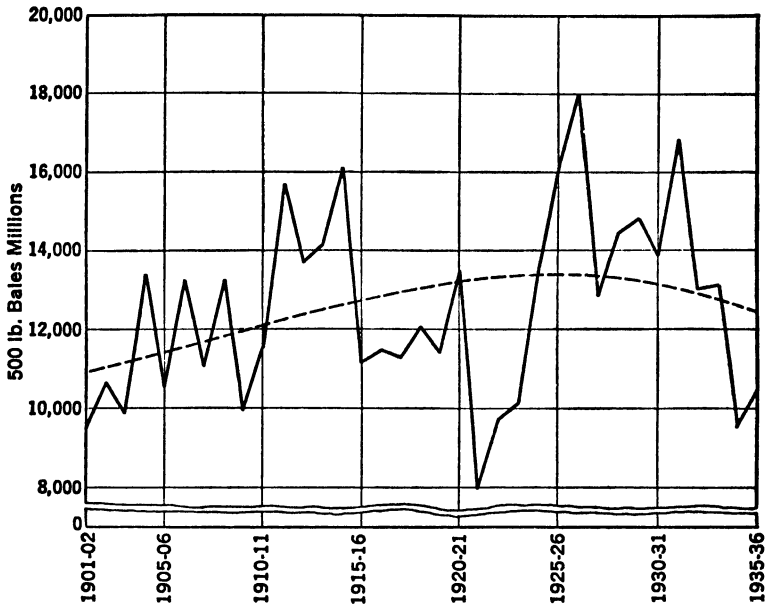


FIG. 76. — Cotton Production in the United States, Crop Years 1901–1902 to 1935–1936, with Line of Trend

starts are average spot prices at New York for middling upland cotton, at wholesale, from September to May of each crop year. But such prices reflect not only the effects of varying conditions in the cotton market, but also changes in the general level of prices. To eliminate the effect of this factor the original prices are deflated by Bradstreet's price index, as computed for the September-May period in each crop year. For this purpose, Bradstreet's index has been reduced to relative terms, with the average for the

TABLE 95

Cotton Production and Cotton Prices, 1901–1936

(1)	(2)	(3)	(4)	(5)
Crop year	*Cotton production in United States, excluding linters (in thousands of bales)*	*Cotton prices. Average of spot prices in N. Y. for middling upland cotton, Sept. to May (in cents per pound)*	*Bradstreet's price index, average, Sept. to May* (1913–14 = 100)	*Cotton prices, deflated (in cents per pound)*
1901–02	9,510	8.64	86.2	10.02
1902–03	10,631	9.50	90.0	10.56
1903–04	9,851	13.20	88.6	14.90
1904–05	13,438	8.69	89.3	9.73
1905–06	10,575	11.40	92.3	12.35
1906–07	13,274	10.97	98.8	11.10
1907–08	11,107	11.41	93.2	12.24
1908–09	13,242	9.81	91.3	10.74
1909–10	10,005	14.62	100.6	14.53
1910–11	11,609	14.80	97.8	15.13
1911–12	15,693	10.34	100.0	10.34
1912–13	13,703	12.35	104.8	11.78
1913–14	14,156	13.40	100.0	13.40
1914–15	16,135	8.63	105.2	8.20
1915–16	11,192	12.04	121.2	9.93
1916–17	11,450	18.29	151.0	12.11
1917–18	11,302	29.96	197.9	15.14
1918–19	12,041	30 06	203.1	14.80
1919–20	11,421	38.63	226.3	17.07
1920–21	13,440	16 90	152.9	11.05
1921–22	7,954	18.67	127.2	14.68
1922–23	9,762	26.26	149.7	17.54
1923–24	10,140	31.79	145.3	21.88
1924–25	13,628	24.34	150.4	16.18
1925–26	16,104	20 60	153.8	13 39
1926–27	17,977	14.26	141.0	10.11
1927–28	12,956	20.19	149.2	13.53
1928–29	14,478	20.02	145.1	13.80
1929–30	14,825	17.00	132.1	12.87
1930–31	13,932	10.47	107.6	9.74
1931–32	17,096	6.42	86.1	7.46
1932–33	13,002	6.75	76.2	8.85
1933–34	13,047	10.95	100.6	10.89
1934–35	9,636	12.42	106.7	11.64
1935–36	10,443	11.59	112.6	10.29

crop year 1913–14 equal to 100. The original figures for the two series to be correlated, together with the corrected price figures, are given in Table 95.

These data are plotted in Figs. 76 and 77. Lines of trend fitted to the two series are shown on the charts.[1]

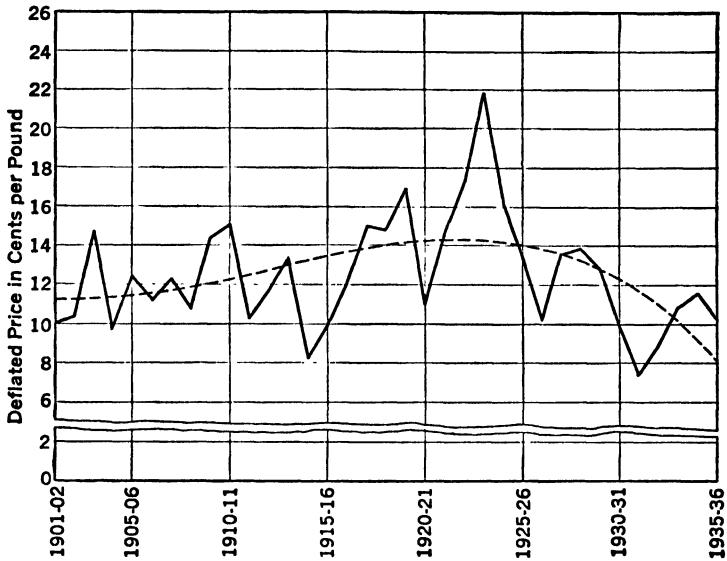


FIG. 77. — Prices of Middling Upland Cotton in New York, Crop Years 1901–1902 to 1935–1936, with Line of Trend. (Figures relate to average annual prices, during crop years, deflated by Bradstreet's index of wholesale prices)

The deviation of each annual item from the secular trend of the given series is now to be measured, and the coefficient of correlation between these deviations is to be calculated. The computations appear in Table 96.

This value of $-.648$ for the coefficient indicates a fair degree of negative correlation between deviations of cotton production in the United States from the line of trend and

[1] The equation to the line of trend of cotton production is

$Y = 13{,}009.14 + 87.96X - 4.640X^2 - .1491X^3$, with origin at 1918–19.

The trend equation for deflated cotton prices is

$Y = 13.96 + .152X - .01425X^2 - .00083X^3$, with origin at 1918–19.

TABLE 96

Computation of Coefficient of Correlation, Cotton Production and Cotton Prices

(1)	(2)	(3)	(4)	(5)	(6)
Crop year	*Deviation of cotton production from trend (in 1,000's of bales)* x	*Deviation of deflated cotton prices from trend (in cents per lb.)* y	x^2	y^2	xy
1901–02	− 1,395	− 1.32	1,946,025	1.7424	+ 1,841.40
1902–03	− 393	− .72	154,449	.5184	+ 282.96
1903–04	− 1,298	+ 3.63	1,684,804	13.1769	− 4,711.74
1904–05	+ 2,161	− 1.59	4,669,921	2.5281	− 3,435.99
1905–06	− 834	+ .95	695,556	.9025	− 792.30
1906–07	+ 1,731	− .42	2,996,361	.1764	− 727.02
1907–08	− 572	+ .57	327,184	.3249	− 326.04
1908–09	+ 1,428	− 1.11	2,039,184	1.2321	− 1,585.08
1909–10	− 1,945	+ 2.49	3,783,025	6.2001	− 4,843.05
1910–11	− 476	+ 2.87	226,576	8.2369	− 1,366.12
1911–12	+ 3,476	− 2.14	12,082,576	4.5796	− 7,438.64
1912–13	+ 1,357	− .93	1,841,449	.8649	− 1,262.01
1913–14	+ 1,648	+ .45	2,715,904	.2025	+ 741.60
1914–15	+ 3,542	− 4.98	12,545,764	24.8004	− 17,639.16
1915–16	− 1,516	− 3.47	2,298,256	12.0409	+ 5,260.52
1916–17	− 1,366	− 1.50	1,865,956	2.2500	+ 2,049.00
1917–18	− 1,615	+ 1.35	2,608,225	1.8225	− 2,180.25
1918–19	− 968	+ .84	937,024	.7056	− 813.12
1919–20	− 1,671	+ 2.97	2,792,241	8.8209	− 4,962.87
1920–21	+ 275	− 3.15	75,625	9.9225	− 866.25
1921–22	− 5,273	+ .41	27,804,529	.1681	− 2,161.93
1922–23	− 3,515	+ 3.25	12,355,225	10.5625	− 11,423.75
1923–24	− 3,174	+ 7.62	10,074,276	58.0644	− 24,185.88
1924–25	+ 290	+ 2.00	84,100	4.0000	+ 580.00
1925–26	+ 2,758	− .65	7,606,564	.4225	− 1,792.70
1926–27	+ 4,637	− 3.73	21,501,769	13.9129	− 17,296.01
1927–28	− 360	− .04	129,600	.0016	+ 14.40
1928–29	+ 1,202	+ .58	1,444,804	.3364	+ 697.16
1929–30	+ 1,608	+ .07	2,585,664	.0049	+ 112.56
1930–31	+ 793	− 2.56	628,849	6.5536	− 2,030.08
1931–32	+ 4,055	− 4.24	16,443,025	17.9776	− 17,193.20
1932–33	+ 80	− 2.17	6,400	4.7089	− 173.60
1933–34	+ 265	+ .66	70,225	.4356	+ 174.90
1934–35	− 2,982	+ 2.30	8,892,324	5.2900	− 6,858.60
1935–36	− 1,988	+ 1.94	3,952,144	3.7636	− 3,856.72
			171,865,603	227.2511	− 128,167.61

TABLE 96—*Continued*

Computation of Coefficient of Correlation, Cotton Production and Cotton Prices

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{171{,}865{,}603}{35}} = 2{,}216.0$$

$$\sigma_y = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{227.2511}{35}} = 2.548$$

$$r = \frac{\Sigma xy}{N\sigma_x \sigma_y} = \frac{-\ 128{,}167.61}{35 \times 2{,}216.0 \times 2.548} = -\ .648.$$

the corresponding deviations of cotton prices in New York, during the period covered.

From the values already computed we may derive an equation for estimating the variation in cotton price associated with a given variation in production. This regression equation, as we have seen, is of the type

$$y = r\frac{\sigma_y}{\sigma_x}x.$$

In the present case y and x refer to deviations from the parabolic lines of trend. Substituting the given values, we have

$$y = -\ .648\frac{2.548}{2{,}216}x$$

$$y = -\ .00074x.$$

This equation means that, on the average, a unit deviation of cotton production (x) above the line of trend was accompanied by a deviation of .00074 units in cotton prices (y) below the line of trend. The unit employed in the production figures was 1,000 bales, in the deflated price figures, one cent. In the interpretation of the equation it may be simpler to use an x-unit of one million bales, making the equation of regression

$$y = -\ .74x.$$

Thus a cotton crop one million bales above trend was

accompanied by prices about three quarters of a cent per pound below trend (with reference always to deflated prices). This was the average relationship during the period 1901–1936. It did not hold in all cases, as is shown by the fact that r has a value of but $-.648$. If this, or a similar law, held perfectly, r would have a value of -1.

The value of S, which measures the scatter about the line of regression, may be computed from the formula

$$S_y = \sigma_y\sqrt{1 - r^2}.$$

In the present case, S_y has a value of 1.94 cents. The significance of this measure has been explained in an earlier section.

(It should be emphasized that the use of the above equation for estimating future prices is dependent upon the validity of projecting the two lines of secular trend.)

In the preceding analysis deviations were measured in absolute units, and the results could be interpreted only in terms of absolute units, bales of cotton and cents per pound. For certain purposes it might have been more convenient to correlate percentage deviations from the two lines of trend, in which case the standard deviations and the equation of regression would have been expressed in these terms. The procedure, in this respect, will depend in part upon the use to which the results are to be put. The nature of the data will also affect a decision on this point. The use of percentage rather than absolute deviations would be desirable in handling series in which the range of absolute deviations had changed materially during the period covered.

It is obvious that in the above problem there is an arbitrary element which was not present in the correlation problems previously studied. The deviations are measured from lines of trend, not from the arithmetic means, and these lines of trend are arbitrarily selected. The use of different lines of trend might give quite different results.

In the above example the lines of trend were both power curves of the third degree. We might, perhaps with equal reason, assume that the underlying trends are best defined by other functions. Coefficients of regression and correlation would have different values if this were done. The presence of this arbitrary element in the correlation of deviations from lines of secular trend detracts somewhat from the confidence that may be placed in the results. The critical problem here lies not in the mechanical process of correlation, but in the choice of an appropriate line of trend for each series. If, by the tests of inspection and of correspondence with such external evidence as may be available, it appears that the curve selected accurately represents the trend in each of the series correlated, the coefficient may be accepted as significant. But, in the interpretation and use of the results, the presence of this element of personal judgment in the preliminary calculations must not be forgotten. This applies with particular force if the study aims to establish a functional relationship between cyclical fluctuations in the two series, and if an estimating (or regression) equation is to be based upon the results.

The Coefficient of Correlation and the Measurement of Time Sequence

In the correlation of cotton production and cotton prices the object was to measure as accurately as possible the effect of variations in cotton production upon cotton prices. An equation was secured which described this relation when deviations were measured from the particular lines of trend employed. Cotton prices were considered to be a function of cotton production, and the object of the study was to measure this functional relationship. We seek, in such cases, to determine the degree to which cycles in one series depend upon or reflect cycles in a related series, assuming some functional relationship between them. This is essentially the problem described in introducing the

subject of correlation, and generally constitutes the major problem in studying the relation between series of any type.

But a second and somewhat different problem may be faced in certain studies of time series. Assuming that two such series are marked by definite cycles, it is of interest to determine whether the cycles coincide in time, or whether cycles in one series consistently precede or lag behind cycles in the other. The coefficient of correlation has been found very useful in determining the degree of "lead" or "lag" in such cases. This problem is that of determining merely *temporal relationship*, as opposed to the *functional relationship* that is ordinarily to be measured.

THE RELATION BETWEEN STOCK PRICE CYCLES AND CYCLES OF BUSINESS ACTIVITY

To illustrate the solution of a problem of this latter type, we may undertake to determine the relation, in time, between cyclical movements in industrial stock prices and in general business activity, as measured by the composite index compiled by the American Telephone and Telegraph Company. The monthly values of this index for the period 1899–1937 have been presented in an earlier section. Figures relating to stock prices from January, 1903, to June, 1914, are given in Table 97.

TABLE 97

Cycles in Industrial Stock Prices, 1903–1914 [1]

(Figures relate to deviations from trend in units of the standard deviation)

Month	1903	1904	1905	1906	1907	1908	1909	1910	1911	1912	1913	1914
January	− .2	− 2.0	− .1	+ 2.3	+ 1 6	− 1.3	+ .5	+ 1.0	− .1	− .4	− .2	− .7
February	− .1	− 2.1	+ .2	+ 2 2	+ 1 4	− 1.5	+ .3	+ .5	+ 1	− .4	− .5	− .6
March	− .3	− 2.1	+ .6	+ 1.9	+ .6	− 1.1	+ .3	+ .8	− .1	− .1	− .7	− .6
April	− .5	− 2.0	+ .7	+ 1.7	+ .6	− .8	+ .5	+ .5	− .1	+ .3	− .6	− .8
May	− .5	− 2.1	+ .2	+ 1.4	+ .4	− .5	+ .8	+ .4	0	+ .2	− .7	− .8
June	− .9	− 2.1	+ .3	+ 1.5	+ .2	− .5	+ .9	+ .1	+ .1	+ .2	− 1.1	− .7
July	− 1.4	− 1.8	+ .7	+ 1.3	+ .3	− .2	+ 1.1	− .4	+ .1	+ .2	− .9	
August	− 1.7	− 1.6	+ .8	+ 1.7	− .3	+ .3	+ 1.4	− .3	− .3	+ .3	− .7	
September	− 1.9	− 1.3	+ .7	+ 1.7	− .5	+ .1	+ 1.4	− .3	− .7	+ .4	− .5	
October	− 2.3	− .9	+ .8	+ 1.7	− 1.3	+ .2	+ 1.4	0	− .7	+ .3	− .8	
November	− 2.4	− .3	+ 1.1	+ 1.7	− 1.9	+ .5	+ 1.4	+ .1	− .5	+ .2	− .9	
December	− 2.1	− .1	+ 1.8	+ 1.7	− 1.6	+ .5	+ 1 3	− .2	− .4	0	− .9	

[1] These figures, the results of analyses by W. M. Persons, are from the *Review of Economic Statistics*, published by the Harvard Committee on Economic Research. They are based upon the average price of 12 industrial stocks.

The data of the two series are plotted in Fig. 78.[1] From a comparison of the two curves in this chart it is clear that there is some relation between the movements in

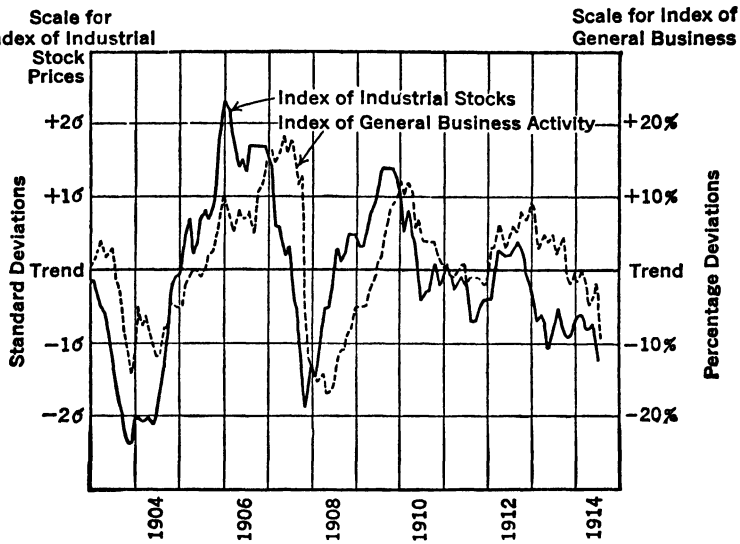


FIG. 78. — Comparison of Cyclical Fluctuations in Industrial Stock Prices and in General Business Activity, 1903–1914

the two series, but such a comparison affords no basis for a definite conclusion. Our object is to determine whether the cycles in the two series are exactly synchronous and, if they are not, to measure the average time interval by which cycles in one series precede the cycles in another. The significance of such studies in the analysis of the business cycle is obvious.

For the study of pre-war relations data for the period from January, 1903, to June, 1914, may be employed. A coefficient of correlation is first computed for concurrent items. A value of + .55 is secured. Next, the data are correlated with industrial stock prices preceding general

[1] The American Telephone and Telegraph index here plotted is not identical with that given in Chapter IX. The latter is a revised series, differing in some respects from the original index for the pre-war period that has been used in the present calculations.

business by one month. That is, the January, 1903, figure for stock prices is multiplied by the February, 1903, index of general business; the February stock price is multiplied by the March business index, etc. This process is carried through for the entire period from January, 1903, to June, 1914. Only 137 monthly values are used in this computation, as compared with 138 in the preceding case, for the January, 1903, business index and the June, 1914, stock price figure do not enter into the calculations. Accordingly, the values c_x and c_y (the two corrections to be applied because the origin does not coincide with the two averages) and the two standard deviations will be slightly different. These corrections may be readily made. The coefficient of correlation secured from these computations has a value of + .65. The same operation is repeated with other pairings of the two variables. The results are summarized below.

TABLE 98

Coefficients of Correlation between Industrial Stock Prices and an Index of General Business Activity

(Based upon data for the period 1903–1914)

	Coefficient of Correlation
Stock prices concurrent with business index	+ .55
Stock prices preceding business index by 1 month	+ .65
" " " " " " 2 months	+ .70
" " " " " " 3 "	+ .73
" " " " " " 4 "	+ .76
" " " " " " 5 "	+ 76
" " " " " " 6 "	+ .76
" " " " " " 7 "	+ .74
" " " " " " 8 "	+ .71
" " " " " " 9 "	+ .67
" " " " " " 10 "	+ .61
" " " " " " 11 "	+ .54

These figures are plotted in Fig. 79.

The coefficients increase to a maximum value of + .76 which is secured with stock prices preceding general business

by 4, 5, and 6 months. The stability of the coefficients with the period of "lead" varying from 3 to 7 months indicates that there was no one specific interval, within the limits thus indicated, between the cyclical movements of these two series. From the results here given it would

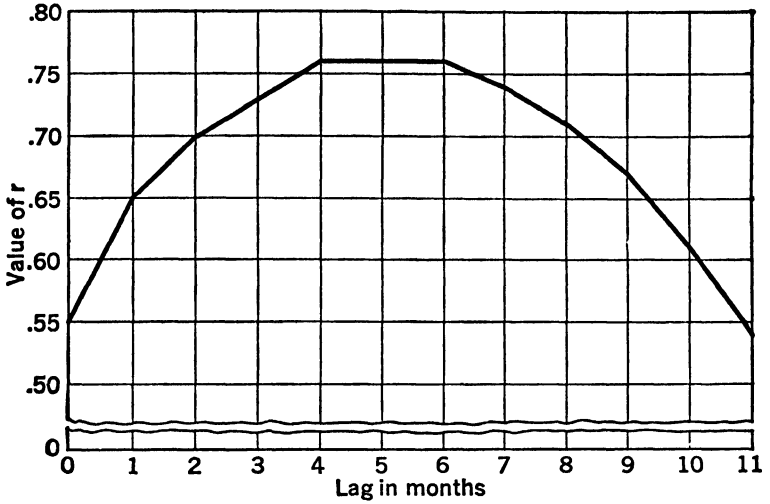


FIG. 79. — Coefficients of Correlation between Index of Industrial Stock Prices and Index of Business Activity, 1903–1914, Showing the Results Secured with Different Pairings. (In all pairings except that of concurrent items the business activity index follows the stock price index)

appear that five months was the average interval by which stock prices preceded the general business index, but this was not sharply marked off as a constant relationship.

With this record of pre-war relations we may contrast the experience of recent years. The Index of Industrial Activity of the American Telephone and Telegraph Company, given in Chapter IX, defines the state of business. Of stock price index numbers, the measurements currently published in the *Review of Economic Statistics*[1] are in a form best

[1] This is not a homogeneous series for the entire period covered. For the years 1919–1924 the index is based on the average price of 20 industrial stocks (the Dow-Jones index), expressed as deviations from trend in units of the standard deviation. For the period 1925–1937 the official all-inclusive index of the New York Stock Exchange (index No. 2) has been used. This index,

(*Footnote 1 continued on page 395*)

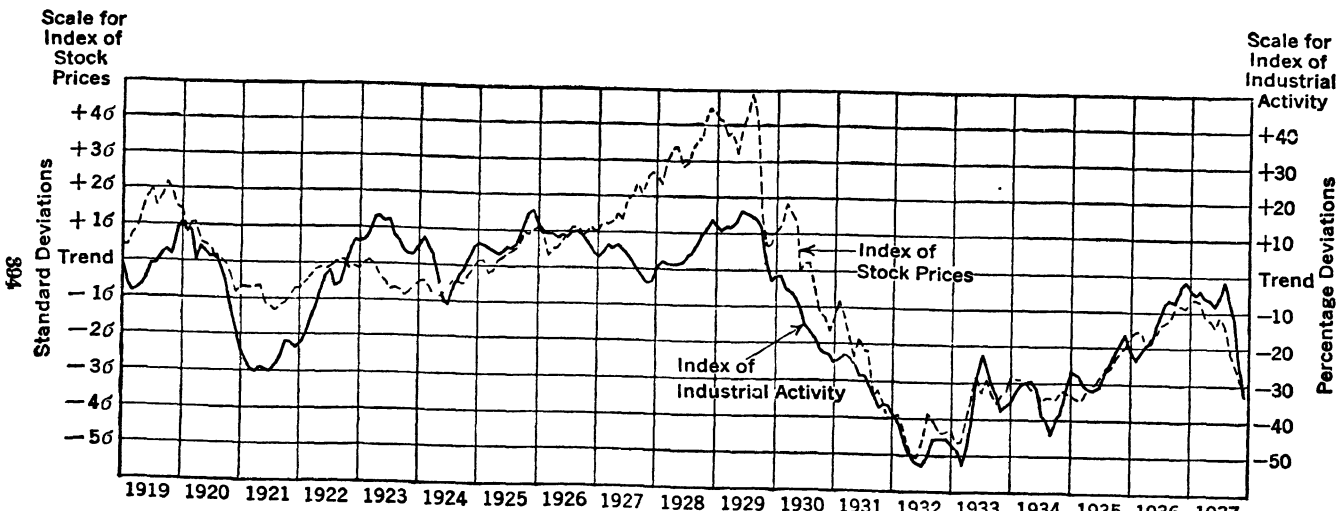


FIG. 80. — Comparison of Cyclical Fluctuations in Stock Prices and Industrial Activity, 1919–1937

adapted to our present needs, although a change in coverage during the period detracts somewhat from their utility for comparative purposes. Monthly values of this index for the period 1919–1937 are recorded in Table 99. The two series are plotted in Fig. 80.

TABLE 99

Cycles in Stock Prices, 1919–1937 [1]

Month	1919	1920	1921	1922	1923	1924	1925	1926	1927	1928
Jan.	+ .38	+1.44	− .69	− .68	− .12	− .47	+ .15	+1.11	+1.04	+2.51
Feb.	+ .38	+ .78	− .70	− .53	+ .05	− 43	+ .17	+ .90	+1.31	+2.31
March	+ .70	+1.06	− .72	− .38	+ 15	− .58	− .21	+ .33	+1.25	+2.96
April	+ .90	+1.10	− .68	− .18	− .02	− .81	− .12	+ .52	+1.32	+3.37
May	+1.40	+ .50	− .67	− .11	− .31	− .88	+ .19	+ .55	+1.56	+3.38
June	+1 76	+ .46	−1.15	− .14	− .46	− .82	+ .27	+ 81	+1 39	+2.82
July	+2.01	+ .38	−1.20	− .08	− .70	− .58	+ .41	+1.00	+1.94	+2.89
Aug.	+1.50	+ .04	−1.32	+ .05	− .65	− .39	+ .43	+1.11	+2.07	+3.37
Sept.	+1.78	+ .11	−1.16	+ .10	− .70	− 44	+ .53	+1.13	+2.42	+3.63
Oct.	+2.14	− .04	−1.12	+ .10	− .84	− .50	+1.01	+ .89	+2.06	+3.69
Nov.	+1.90	− .44	− .89	− .16	− .72	− .25	+ .90	+1.07	+2.55	+4.39
Dec.	+1.54	− .85	− .70	− 10	− .60	− .02	+1.05	+1.05	+2.67	+4.41

Month	1929	1930	1931	1932	1933	1934	1935	1936	1937
Jan.	+4.21	+1.11	−1.35	−4 02	−4.31	−2.83	−3 30	−1.61	− .64
Feb.	+4.11	+1.28	− .83	−3.90	−4.65	−2 90	−3.37	−1.51	− .61
March	+3.70	+1.83	−1 22	−4.20	−4.62	−2.89	−3.51	−1.51	− .65
April	+3 84	+1.59	−1.73	−4 64	−3 91	−2.93	−3.23	−1.92	−1.11
May	+3 17	+1.41	−2 35	−5.05	−3.33	−3.19	−3.14	−1.71	−1.17
June	+3.88	+ .13	−1 85	−5.09	−2 90	−3.13	−2.97	−1.61	−1.44
July	+4.25	+ .25	−2 15	−4.60	−3.26	−3.51	−2.71	−1.31	−1.03
Aug.	+4.89	+ .24	−2.18	−3.86	−2.89	−3.37	−2.62	−1.27	−1.31
Sept.	+4.10	− .51	−3.42	−3.97	−3 31	−3.40	−2 55	−1.23	−2.04
Oct.	+1.67	−1 04	−3.23	−4.30	−3.57	−3.45	−2.29	− .90	−2.48
Nov.	+ .68	−1.21	−3.54	−4.42	−3.32	−3.21	−2.10	− .78	−2.85
Dec.	+ .74	−1.65	−3.99	−4.37	−3.27	−3.21	−1.93	− .81	−3.03

Results obtained from a study of the temporal relations between these two series, for the period 1919–1937, are given in Table 100.

(*Footnote 1 continued from page 393*)
originally constructed with the figure for Jan. 1, 1925 as 100, has here been expressed in terms of deviations from 100, in units of a standard deviation assumed to be equal to 15 on the original scale. In effect, a horizontal trend at the level of Jan. 1, 1925, has been assumed for the Stock Exchange index. This index has also been shifted slightly in time. The index figure relating to the first day of a given month, in the Stock Exchange tabulations, has here been recorded as for the month preceding. Thus a February 1st index is entered for January, a March 1st index for February, etc.

[1] From the *Review of Economic Statistics*. The figures in the table define deviations from trend, in units of the standard deviation, with the assumptions stated in the preceding footnote. The coefficients in Table 100 are based upon data through July, 1937, only.

TABLE 100

Coefficients of Correlation between Stock Prices and an Index of Business Activity

(Based upon data for the period 1919–1937)

	Coefficient of Correlation
Stock prices concurrent with business index	+ .85
Stock prices preceding business index by 1 month	+ .86
" " " " " " 2 months	+ .85
" " " " " " 3 "	+ .83
" " " " " " 4 "	+ .82
" " " " " " 5 "	+ .80
" " " " " " 6 "	+ .78
" " " " " " 7 "	+ .76
" " " " " " 8 "	+ .74
" " " " " " 9 "	+ .71
" " " " " " 10 "	+ .68
" " " " " " 11 "	+ .65

These measurements are shown graphically in Fig. 81.

In using these coefficients we should note that the stock price records for part of the recent period are different in important respects from those employed for the pre-war period. In place of the 12 industrial stocks entering into the earlier comparisons the index for the recent period included 20 stocks and, later, a comprehensive list composed of all varieties of stocks. The market behavior of the broader list may have departed somewhat from the pattern set by the limited number of industrial stocks. The difference between the results for the two periods is to be interpreted with this fact in mind.

In post-war years the highest degree of correlation prevailed with the business index following the stock price index by one month. The traditional "lead" of stock prices, on the basis of which the movements of these prices have been used as forecasters of business changes, was clearly reduced in this period. The actual statistical record we have obtained may have been affected somewhat by the broadening of the coverage of the stock price index

used, but the change in the relations between the two series appears to have been a real one.

This method of measuring temporal relations between economic series is highly useful, but one important caution should be noted. The method indicates the *average* degree of lead or lag of one series, with reference to another. Frequently the sequences of change in economic series are not the same in all phases of business cycles. Thus, observations relating to ten business cycles occurring between 1890

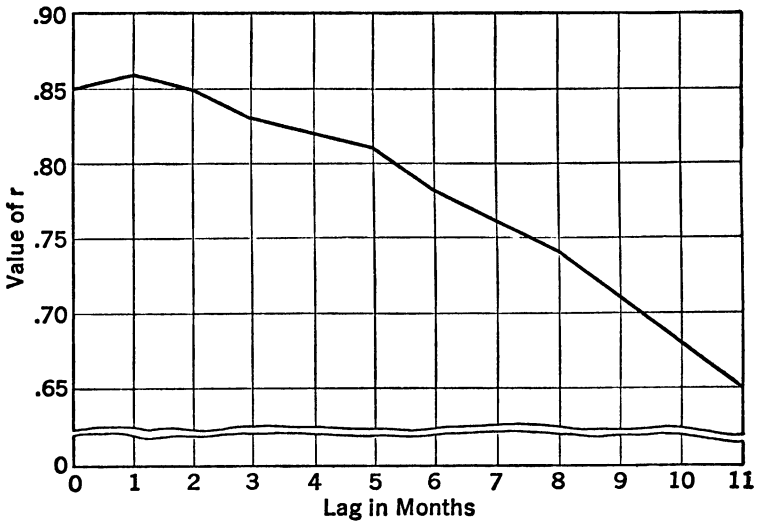


FIG. 81. — Coefficients of Correlation between Index of Industrial Stock Prices and Index of Business Activity, 1919–1937, Showing the Results Secured with Different Pairings. (In all pairings except that of concurrent items the business activity index follows the stock price index)

and 1925 indicate that pig iron prices *preceded* the general index of wholesale prices by 3.4 months, on the average, in business recessions, but *followed* the general index by 5.1 months, on the average, in periods of business revival.[1] This highly important difference would be ironed out in the measurement of average temporal relations by the

[1] Cf. *The Behavior of Prices*, New York, National Bureau of Economic Research, 1927, 84–87.

correlation method. The use of an average should be supplemented by a study of the items entering into the average. Study of the relations among individual observations at different cyclical phases is essential when correlation technique is employed to define time sequences among the movements of economic series.

The Use of the Moving Average in Correlating Cycles in Time Series

The preceding discussion has dealt only with cycles as measured from mathematically fitted lines of trend. But trend may be measured, as we have seen, by lines based upon moving averages, and the cyclical deviations from such lines may be correlated in precisely the same way as deviations from other lines of trend. The arithmetic mean of the deviations from such moving averages will not necessarily be zero, as in the case of deviations measured from lines fitted by the method of least squares, and a corresponding correction must be made in correlating such figures.

Moving averages are subject to the same criticism as are mathematical lines of trend. There can be no certainty that deviations from lines of trend based upon moving averages represent the effects of cyclical causes solely. The result in a given case depends upon the period of the moving average employed, and there is no perfect criterion by which to determine the best measure of trend. Significant and useful coefficients may be computed when deviations are measured from moving averages, but the presence of an arbitrary element in the work must be recognized and the results applied with corresponding reservations.

The Correlation of Short Term Fluctuations

In describing the variable factors that constitute component elements of the values of a series in time, it was pointed out that the coefficient of correlation would not generally be employed in comparing either the secular trends or the

seasonal fluctuations of two series. It may be used to advantage in measuring either *functional* or *temporal* relations between cyclical fluctuations, provided that the effects of the other variables have been, so far as possible, eliminated. The coefficient of correlation and the measures which are employed in conjunction with it have a further use in dealing with time series. They may be used to measure the relation between short term changes in two series, changes from year to year, month to month, or even from week to week or day to day, if desired. This problem is distinct from that studied in the preceding section and in the interpretation of the results the two should not be confused.

There are several ways in which the problem of comparing short term fluctuations may be attacked. The absolute differences between successive items in two series may be correlated, or these differences may be expressed as percentages or ratios. Table 101 illustrates the procedure employed in measuring the correlation between the absolute fluctuations from year to year (first differences) of cotton production and cotton prices. The original values from which the items in columns (2) and (3) are derived are given in Table 95.

The process of computing r is identical with that employed in preceding examples, when deviations were measured from an arbitrary origin. The arbitrary origin in this case is zero, but corrections must be made in the various values since the algebraic sum of the given figures is not zero in either case. Computations based on the figures in Table 101 follow:

$$c_x = \frac{\Sigma X}{N} = \frac{+\ .933}{34} = +\ .02744$$

$$c_x^2 = .000753$$

$$\sigma_x = \sqrt{\frac{\Sigma X^2}{N} - c_x^2} = \sqrt{\frac{229.624987}{34} - .000753} = 2.599$$

TABLE 101

Computation of Coefficient of Correlation between Cotton Production and Cotton Prices, 1902–1936

(Based upon first differences)

(1)	(2)	(3)	(4)	(5)	(6)
Crop Year	*Difference between production in given year and production in preceding year (in millions of bales)* X	*Difference between price in given year and price in preceding year (in cents per pound, deflated)* Y	X^2	Y^2	XY
1902–03	+ 1.121	+ .54	1.256641	.2916	+ .60534
1903–04	− .780	+ 4.34	.608400	18.8356	− 3.38520
1904–05	+ 3.587	− 5.17	12.866569	26.7289	− 18.54479
1905–06	− 2.863	+ 2.62	8.196769	6.8644	− 7.50106
1906–07	+ 2.699	− 1.25	7.284601	1.5625	− 3.37375
1907–08	− 2.167	+ 1.14	4.695889	1.2996	− 2.47038
1908–09	+ 2.135	− 1.50	4.558225	2.2500	− 3.20250
1909–10	− 3.237	+ 3.79	10.478169	14.3641	− 12.26823
1910–11	+ 1.604	+ .60	2.572816	.3600	+ .96240
1911–12	+ 4.084	− 4.79	16.679056	22.9441	− 19.56236
1912–13	− 1.990	+ 1.44	3.960100	2.0736	− 2.86560
1913–14	+ .453	+ 1.62	.205209	2.6244	+ .73386
1914–15	+ 1.979	− 5.20	3.916441	27.0400	− 10.29080
1915–16	− 4.943	+ 1.73	24.433249	2.9929	− 8.55139
1916–17	+ .258	+ 2.18	.066564	4.7524	+ .56244
1917–18	− .148	+ 3.03	.021904	9.1809	− .44844
1918–19	+ .739	− .34	.546121	.1156	− .25126
1919–20	− .620	+ 2.27	.384400	5.1529	− 1.40740
1920–21	+ 2.019	− 6.02	4.076361	36.2404	− 12.15438
1921–22	− 5.486	+ 3.63	30.096196	13.1769	− 19.91418
1922–23	+ 1.808	+ 2.86	3.268864	8.1796	+ 5.17088
1923–24	+ .378	+ 4.34	.142884	18.8356	+ 1.64052
1924–25	+ 3.488	− 5.70	12.166144	32.4900	− 19.88160
1925–26	+ 2.476	− 2.79	6.130576	7.7841	− 6.90804
1926–27	+ 1.873	− 3.28	3.508129	10.7584	− 6.14344
1927–28	− 5.021	+ 3.42	25.210441	11.6964	− 17.17182
1928–29	+ 1.522	+ .27	2.316484	.0729	+ .41094
1929–30	+ .347	− .93	.120409	.8649	− .32271
1930–31	− .893	− 3.13	.797449	9.7969	+ 2.79509
1931–32	+ 3.164	− 2.28	10.010896	5.1984	− 7.21392
1932–33	− 4.094	+ 1.39	16.760836	1.9321	− 5.69066
1933–34	+ .045	+ 2.04	.002025	4.1616	+ .09180
1934–35	− 3.411	+ .75	11.634921	.5625	− 2.55825
1935–36	+ .807	− 1.35	.651249	1.8225	− 1.08945
	+ .933	+ .27	229.624987	313.0067	−180.19834

$$c_y = \frac{\Sigma Y}{N} = \frac{+ .27}{34} = + .00794$$

$$c_y^2 = .000063$$

$$\sigma_y = \sqrt{\frac{\Sigma Y^2}{N} - c_y^2} = \sqrt{\frac{313.0067}{34} - .000063} = 3.034$$

$$p = \frac{\Sigma(XY)}{N} - c_x c_y = \frac{- 180.19834}{34} - (.02744 \times .00794)$$

$$p = - 5.300168$$

$$r = \frac{p}{\sigma_x \sigma_y} = \frac{- 5.300168}{2.599 \times 3.034}$$

$$r = - .672.$$

The equation of regression and the value of S_y, computed from the usual formulas, are

$$y = - .78x$$
$$S_y = 2.25 \text{ cents.}$$

A comparison of the different results secured in the preceding examples relating to cotton throws some interesting light upon the general problem of correlation. In fact, in the two examples, we have measured the correlation between measurements that are not strictly comparable — deviations from third degree parabolas, in the first case, and year-to-year fluctuations in the production and price of cotton, in the second. Yet, if we were seeking to estimate the price of cotton which would accompany a given crop, an estimate might be based upon either of the studies, the results of which are given below.

	r	S_y
Correlation of cycles in cotton production and prices (deviations measured from third degree parabolas)	— .648	1.94 cents
Correlation of year-to-year fluctuations, same data	— .672	2.25 cents

The value of r in the second example is slightly greater than the value secured in the first case, though the standard error is also larger. The reason for this apparent contradiction has been suggested above; the standard deviation of

the year-to-year fluctuations in cotton prices is greater than the standard deviation about the trend of cotton prices.

It appears that errors of estimate are less when based upon the results secured when deviations from third degree curves are correlated than when based upon the study of year-to-year movements. But there is a concealed assumption in the first case, the assumption that the lines of trend of both prices and production may be projected beyond the period studied. There is an immeasurable margin of error in this assumption, and the standard error of estimate, accordingly, does not give a true measure of the probabilities involved. No such assumption is involved in the measure based upon year-to-year fluctuations.

REFERENCES

Boddington, A. L., *Statistics and Their Application to Commerce,* Chap. 10.

Crum, W. L. and Patton, A. C., *An Introduction to the Methods of Statistics,* Chap. 23.

Day, E. E., *Statistical Analysis,* Chap. 20.

Jerome, Harry, *Statistical Method,* Chap. 25.

Kuznets, Simon, "On Moving Correlation of Time Sequences," *Journal of the American Statistical Association,* June, 1928.

Mitchell, W. C., *Business Cycles,* Chap. 3.

Moore, H. L., *Economic Cycles: Their Law and Cause.*

Moore, H. L., *Forecasting the Yield and the Price of Cotton.*

Moore, H. L., *Generating Economic Cycles.*

Persons, W. M., "Correlation of Time Series." *Journal of the American Statistical Association,* June, 1923. (This article is also published in *Handbook of Mathematical Statistics,* Rietz, H. L., ed., Chap. 10.)

Persons, W. M., "Indices of Business Conditions." *Review of Economic Statistics,* Prel. Vol. I. 1919.

Persons, W. M., "The Variate Difference Correlation Method and Curve Fitting." *Quarterly Publications of the American Statistical Association,* June, 1917.

Smith, B. B., "Combining the Advantages of First-Difference and Deviation-from-Trend Methods of Correlating Time Series." *Journal of the American Statistical Association,* Vol. 21, 1926.

Snow, E. C., "Trade Forecasting and Prices." *Journal of the Royal Statistical Society*, May, 1923 (332–398)

Stamp, J. C., "The Effect of Trade Fluctuations Upon Profits." *Journal of the Royal Statistical Society*, July, 1918 (563–608).

Yule, G. U., "On the Time Correlation Problem, with Especial Reference to the Variate Difference Correlation Method." *Journal of the Royal Statistical Society*, July, 1921 (497–537).

Yule, G. U., "Why do we Sometimes get Nonsense Correlations Between Time Series? A Study in Sampling and the Nature of Time Series." *Journal of the Royal Statistical Society*, Vol. 89, 1926.

CHAPTER XII

THE MEASUREMENT OF RELATIONSHIP: NON-LINEAR CORRELATION

In the preceding chapters the discussion has been confined to cases in which the relationship between two variables may be described by a straight line. The coefficient of correlation, r, is a measure of the degree to which two variables approach a linear relationship and it is significant only when a straight line gives a good fit to the points representing the paired values of X and Y.

In fitting curves to time series, as explained in an earlier section, it is found that in many cases the trend is non-linear, and that a curve of higher degree is needed. The same thing is true in the field of our present discussion. It is possible to have a high degree of correlation between two variables when a straight line does not describe the relationship. In such a case there would be considerable scatter about the straight line of best fit, and the value of r would be misleadingly low. If a curve representing the real relationship could be fitted, the scatter would be materially reduced and the true correlation could be measured. The figures presented in Table 102 illustrate such a case. These data are plotted in Fig. 82.

Two different curves have been fitted to the points plotted in this figure. One is a straight line having the equation

$$Y = 5.038 + .0886X$$

in which Y represents yield, in tons per acre, and X represents depth of irrigation water applied, in inches. The degree of relationship between the two variables, as de-

scribed by this line, is indicated by the coefficient of correlation, r, which has a value of + .69.

TABLE 102

Alfalfa Yield and Irrigation

Summary of investigations at Davis, California [1]

(The measurements in the body of the table measure yields, in tons per acre, in 44 experiments)

	Inches of irrigation water applied								
	0	12	18	24	30	36	48	60	
	2.35	4.31	5.69	6.00	7.53	7.58	8.05	5.55	
	2.75	4.78	6.46	6.89	7.97	8.22	8.45	7.25	
	2.89	4.84	7.02	7.96	8.32	8.63	8.63	10.17	
	3.85	5.83	8.02	8.32	9.43	9.33	8.83	10.70	
	5.52	6.51		8.38	9.54	9.38	9.52		
Average	5.94	7.52		9.96	11.06	12.48	10.62		
yield	3.88	5.63	6.80	7.92	8.98	9.27	9.02	8.42	7.48

An inspection of the figure shows clearly that the straight line does not give the best possible fit. It is certain, therefore, that r does not furnish a valid measure of the degree of relationship between alfalfa yield and depth of irrigation water.

PARABOLIC RELATIONSHIP

The other curve in Fig. 82 is a second degree parabola, fitted by the method of least squares. The equation to this curve is

$$Y = 3.539 + .2527X - .002827X^2.$$

It is obvious that the effect of increasing irrigation upon alfalfa yield is described much more accurately by this latter curve than by the straight line. The most important result of these investigations was the determination of the point at which alfalfa yield began to fall off with increased applications of water, and the straight line fails to indicate any such decline.

[1] This table is taken from "The Economical Irrigation of Alfalfa in the Sacramento Valley" by S. H. Beckett and R. D. Robertson, *Bull. No. 280*, Agricultural Experiment Station, Univ. of California, May, 1917.

As the equation of relationship, therefore, we should use the parabolic rather than the linear form. The standard error, S_y, which is a necessary accompanying measure, may be calculated by measuring the deviation of each value from the corresponding computed value, and determining

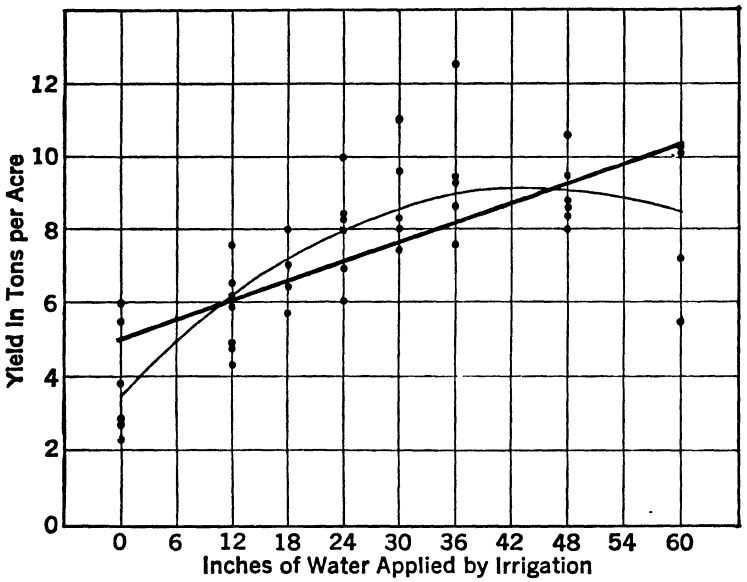


FIG. 82. — Scatter Diagram Showing the Relation between Alfalfa Yield and Irrigation Water Applied, with Two Lines of Regression

the root-mean-square of these deviations. This procedure is illustrated in Table 103. The figures for normal yield which are given in this table are computed from the parabolic equation given above.

Inserting the sum of the squared deviations, as given in col. (5) of Table 103, in the formula

$$S_y = \sqrt{\frac{\Sigma d^2}{N}}$$

we have

$$S_y = \sqrt{\frac{80.9945}{44}} = 1.36.$$

TABLE 103

Comparison of Actual and Computed Alfalfa Yield

(1) *Depth of irrigation water* X	(2) *Actual yield* Y	(3) *Normal yield, as computed from parabolic equation* Y_c	(4) *Deviation of actual from normal* (2) − (3) d	(5) d^2
0	3.85	3.54	+ .31	.0961
0	5.94	3.54	+ 2.40	5.7600
0	5.52	3.54	+ 1.98	3.9204
0	2.75	3.54	− .79	.6241
0	2.89	3.54	− .65	.4225
0	2.35	3.54	− 1.19	1.4161
12	4.78	6.16	− 1.38	1.9044
12	7.52	6.16	+ 1.36	1.8496
12	6.51	6.16	+ .35	.1225
12	4.31	6.16	− 1.85	3.4225
12	5.83	6.16	− .33	.1089
12	4.84	6.16	− 1.32	1.7424
18	7.02	7.17	− .15	.0225
18	5.69	7.17	− 1.48	2.1904
18	8.02	7.17	+ .85	.7225
18	6.46	7.17	− .71	.5041
24	6.00	7.98	− 1.98	3.9204
24	8.38	7.98	+ .40	.1600
24	8.32	7.98	+ .34	.1156
24	6.89	7.98	− 1.09	1.1881
24	9.96	7.98	+ 1.98	3.9204
24	7.96	7.98	− .02	.0004
30	7.53	8.58	− 1.05	1.1025
30	9.54	8.58	+ .96	.9216
30	9.43	8.58	+ .85	.7225
30	7.97	8.58	− .61	.3721
30	11.06	8.58	+ 2.48	6.1504
30	8.32	8.58	− .26	.0676
36	7.58	8.97	− 1.39	1.9321
36	9.33	8.97	+ .36	.1296
36	9.38	8.97	+ .41	.1681
36	8.22	8.97	− .75	.5625
36	12.48	8.97	+ 3.51	12.3201
36	8.63	8.97	− .34	.1156
48	8.45	9.16	− .71	.5041
48	9.52	9.16	+ .36	.1296

(*Continued on next page*)

TABLE 103 (*Continued*)

Comparison of Actual and Computed Alfalfa Yield

(1)	(2)	(3)	(4)	(5)
Depth of irrigation water	*Actual yield*	*Normal yield as computed from parabolic equation*	*Deviation of actual from normal* (2) − (3)	
X	Y	Y_c	d	d^2
48	8.63	9.16	− .53	.2809
48	8.83	9.16	− .33	.1089
48	10.62	9.16	+ 1.46	2.1316
48	8.05	9.16	− 1.11	1.2321
60	10.17	8.52	+ 1.65	2.7225
60	7.25	8.52	− 1.27	1.6129
60	10.70	8.52	+ 2.18	4.7524
60	5.55	8.52	− 2.97	8.8209
				80.9945

THE INDEX OF CORRELATION

We need now the third value, the abstract measure of degree of relationship. In dealing with cases of linear relationship in the preceding chapter we found that such a measure, the coefficient of correlation, could be derived from known values of S_y and σ_y. An analogous measure may be derived in the same way in cases of non-linear relationship, such as that found in the present problem. Since the term *coefficient of correlation* and the symbol r refer only to cases of linear regression, we may term this general measure the *index of correlation*, and use the symbol ρ (rho) to represent it.

As a general formula for the index of correlation we have[1]

[1] With X dependent this formula becomes

$$\rho_{xy}^2 = 1 - \frac{S_x^2}{\sigma_x^2}.$$

The first of the two subscripts refers always to the dependent variable, the second to the independent. It is essential that these be shown, for ρ would not necessarily be the same with X dependent as with Y dependent. Such a distinction is not necessary in the case of linear correlation, for r is the same no matter which variable be dependent.

$$\rho_{yx}^2 = 1 - \frac{S_y^2}{\sigma_y^2}.$$

The value of S_y has been derived above. The value of σ_y, computed by familiar methods, is found to be 2.27. Substituting in the formula for ρ, we have

$$\rho_{yx} = \sqrt{1 - \frac{1.84}{5.19}}$$
$$= .80.$$

This value is materially greater than that of the coefficient of correlation for the same data. The value of r is + .69. The difference is due to the fact that the second degree parabola constitutes a much better fit to the data than the straight line. The correlation is distinctly non-linear, and r is an inappropriate measure of correlation.

THE MEANING OF THE INDEX OF CORRELATION

It is important that the significance and the limitations of ρ be understood. Its value depends upon the relation between the scatter about the fitted line and the scatter about the arithmetic mean of the Y's. In the case of a straight line, ρ and r are identical, r being a special case of ρ. The limits of ρ are 0 and 1, a value of 0 indicating that there is no relationship, or that if there is a relationship between the two variables it cannot be described by the particular equation employed. A value of 1 indicates that the relationship, as described by the equation employed, is a perfect one. For curves of higher degree no positive or negative sign should be attached to ρ, for the relationship might be positive over part of the range and negative over other parts, as in the alfalfa example given above.

The index of correlation, ρ, has no significance unless the type of curve to which it applies be named in each case. The meaning of r in this respect is always clear, for it is understood that it relates always to a straight line, but confusion would arise in the case of ρ unless the type of

curve were specifically mentioned. The index of correlation may be looked upon as a measure of the adequacy of a curve of given type to describe the relationship between two variables.

It is, of course, always possible to secure a curve which will pass through any number of points if the constants in the equation be equal to the number of points. In such a case ρ would, of necessity, be equal to 1, but this value would have no significance. In any employment of mathematical functions there is this limit of absurdity, when the number of constants is equal to the number of points, and ρ would merely reflect this absurdity. The ordinary principles of curve fitting must be kept in mind in using such an index as this. It must never be taken to have an absolute significance, standing by itself. Its significance is always relative, referring to the particular function employed. This fact, which is true of every measure of correlation, is frequently overlooked, and invalid and fallacious conclusions reached as a result.

A SHORT METHOD OF COMPUTING THE INDEX OF CORRELATION

The standard error and the index of correlation were computed by a rather laborious method in the above example, in order that there might be no misunderstanding of their precise meaning. The burden of calculation may be materially reduced, however, by taking advantage of the relationships which were disclosed in dealing with r. For a curve of the potential series

$$Y = a + bX + cX^2 + dX^3 \ldots$$

the formula for S_y is derived by a simple extension of that employed in the case of the straight line. As a general formula for a series of this type, we have

$$S_y^2 = \frac{\Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY) - c\Sigma(X^2Y) - d\Sigma(X^3Y) - \ldots}{N}.$$

Similarly, the formula for r may be extended to give a

general formula for ρ applicable to any equation of this general type. This formula[1] is

$$\rho_{yx}^2 = \frac{a\Sigma(Y) + b\Sigma(XY) + c\Sigma(X^2Y) + d\Sigma(X^3Y) + \ldots - Nc_y^2}{\Sigma(Y^2) - Nc_y^2}.$$

In the special case in which the origin is at the mean of the Y's, $\Sigma(y) = 0$ and $c_y = 0$, and the formula reduces to

$$\rho_{yx}^2 = \frac{b\Sigma(Xy) + c\Sigma(X^2y) + d\Sigma(X^3y) + \ldots}{\Sigma(y^2)}.$$

The characteristics of the formulas for S and ρ should be noted. The only values required in securing these measures are the constants in the equation which describes the average relationship, certain values which have been used in the process of fitting and, in addition, $\Sigma(Y^2)$ and c_y^2. Thus, as direct by-products of the fitting process, we have the values of S and ρ, the two measures which are needed to supplement the regression equation in securing a complete description of the relationship between the two variables in question. The equation describes the average relationship. The standard error, S, is a measure of the reliability of estimates based upon this equation, and ρ is an abstract index of the degree of relationship, in so far as that relationship can be described by the particular curve employed.

The application of these formulas may be illustrated with reference to the problem of alfalfa yield. The following values, derived from the data of Table 102 and from the fitting process, are required for this purpose:

$a = 3.539$	$\Sigma(X^2Y) = 407{,}564.64$
$b = .252652$	$c_y^2 = 55.9197$
$c = -.002827$	$\Sigma(Y^2) = 2{,}688.2268$
$\Sigma(Y) = 329.03$	$N = 44.$
$\Sigma(XY) = 10{,}271.72$	

Substituting in the formula for the standard error for a

[1] See Appendix A for a discussion of the derivation of this formula.

second degree parabola,

$$S_y^2 = \frac{\Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY) - c\Sigma(X^2Y)}{N},$$

we have

$$S_y^2 = \frac{2{,}688.2268 - (3.539 \times 329.03) - (.252652 \times 10{,}271.72) - (-.002827 \times 407{,}564.64)}{44}$$

$$= \frac{80.8043}{44}$$

$$= 1.8365$$

$$S_y = 1.36.$$

The index of correlation, for a curve of this type, is computed from the equation

$$\rho_{yx}^2 = \frac{a\Sigma(Y) + b\Sigma(XY) + c\Sigma(X^2Y) - Nc_y^2}{\Sigma(Y)^2 - Nc_y^2}.$$

Substituting the appropriate values, we have

$$\rho_{yx}^2 = \frac{146.9557}{2{,}688.2268 - (44 \times 55.9197)}$$

$$= .6452$$

$$\rho_{yx} = .80.$$

The value of the index of correlation is influenced by the relation between the number of observations and the number of constants in the equation of relationship. When the two are equal ρ will have a value of 1. In any case the observed index of correlation tends to exceed the true index. When the number of observations is not large it is advisable to apply a correction for this bias. If we use $\bar{\rho}$ to represent the corrected value and m to represent the number of constants in the equation of relationship, we may apply a correction in terms of the relation[1]

$$\bar{\rho}_{yx}^2 = 1 - \left\{ (1 - \rho_{yx}^2)\left(\frac{N-1}{N-m}\right) \right\}.$$

Inserting the values given in the above example, we have

[1] From Mordecai Ezekiel, *Methods of Correlation Analysis*, New York, Wiley, 1930, 121.

$$\bar{\rho}_{yx}^2 = 1 - \left\{(1 - .6452)\left(\frac{44 - 1}{44 - 3}\right)\right\}$$
$$= .6279$$
$$\bar{\rho}_{yx} = .79.$$

If, in the application of this test, the value in brackets { } exceeds unity, the value of $\bar{\rho}$ is taken as 0.[1]

These methods of deriving S and ρ are applicable over a wide field by a simple adaptation of the formulas to the particular equations that may be employed in given instances. Further illustrations are given in Chapter XVII, while this general method is explained in more detail in Appendix A.

The Correlation Ratio

A third distinctive measure of correlation remains to be described. This is the *correlation ratio*, devised by Karl Pearson and represented by the symbol η (eta). This measure may be looked upon as a special case of ρ, but somewhat different methods are employed in its computation.

We have seen that in all cases the degree of relationship between two variables, as described by a curve of a given type, may be determined from the formula

$$\text{Measure of correlation} = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}.$$

The coefficient of correlation, r, is just such a measure, when S_y represents the standard deviation about a straight line. The index of correlation, ρ, is a general measure of the same type. The *correlation ratio* is precisely the same sort of measure, S_y in this case representing the standard deviation about a line passing through the mean of every

[1] A corresponding correction should be made in the standard error of estimate, when derived from a small number of observations. In this case the correction must raise the unadjusted measure. For this correction Ezekiel gives

$$\overline{S^2} = S^2\left(\frac{N - 1}{N - m}\right)$$

where $\overline{S}$ represents the corrected standard error of estimate.

column in the correlation table. We have, in effect, increased the number of constants in the equation of the curve to be fitted until the number is equal to the number of columns. If the means of all the columns lie on a straight line, the correlation ratio and the coefficient of correlation will be equal. If the means of the columns do not lie on a straight line, the correlation ratio will be greater than the coefficient of correlation.

No new principle is involved, therefore, in the concept of the correlation ratio. It is employed when the regression is non-linear. It measures the degree of relationship between two variables, in so far as this relationship may be described by a curve passing through the mean of every column. If the relationship is perfect, if there is no scatter about the curve fitted in this way, η will have a value of 1. If there is no relationship, if the scatter about the curve is as great as the dispersion about the mean of the Y's, η will have a value of zero.

The formula generally employed in the computation of the correlation ratio differs somewhat from that given above. To represent the standard deviation about the line joining the means of the columns, the symbol σ_{ay} is employed, instead of S_y. Its meaning is precisely the same as that of S_y, as employed above, except that σ_{ay} refers always to a correlation table.

The formula may be written

$$\eta_{yx} = \sqrt{1 - \frac{\sigma_{ay}^2}{\sigma_y^2}}.$$

When *eta* is written as above (η_{yx}) it refers to the regression of Y on X (Y dependent). When it is written η_{xy} it refers to the regression of X on Y (X dependent), and its value depends upon the scatter about a line joining the means of the rows. Unlike r, which has the same value for both regressions, η_{yx} and η_{xy} will have different values unless the regression be linear.

THE COMPUTATION OF THE CORRELATION RATIO

Table 104 shows the general relation between the amount of nitrogen, in pounds per acre, used as fertilizer in certain agricultural experiments, and the corresponding yield of wheat, in bushels per acre.[1] The points are plotted in Fig. 83.

TABLE 104

Correlation Table Showing the Relation between Wheat Yield per Acre and Amount of Nitrogen Used as Fertilizer

		X — Nitrogen applied in pounds per acre										
		0–19.9	20–39.9	40–59.9	60–79.9	80–99.9	100–119.9	120–139.9	140–159.9	160–179.9	Total	Mean of Rows
	32–35.9				5	16	12	4	5	2	44	107.27
Y — Wheat yield in bushels per acre	28–31.9			1	20	21	8	4	1		55	88.91
	24–27.9			16	19						35	60.86
	20–23.9			13							13	50.0
	16–19.9		12								12	30.0
	12–15.9		8								8	30.0
	8–11.9	3	5								8	22.50
	4–7.9	10									10	10.0
	0–3.9	8									8	10.0
	Total	21	25	30	44	37	20	8	6	2	193	
	Mean of Columns	5.05	15.12	24.4	28.73	31.73	32.4	32.0	33.33	34.0		

[1] This table is based upon experiments described by E. Davenport ("Comparative Agriculture" in Bailey's *Cyclopedia of American Agriculture*). The

For the computation of η_{yx} by the formula given above we need the values of σ_y and σ_{ay}, the latter being the root-mean-square deviation about the line joining the means of the various columns. The former value may be obtained readily by methods already familiar. It is possible to compute the quantity σ_{ay} by the method first employed

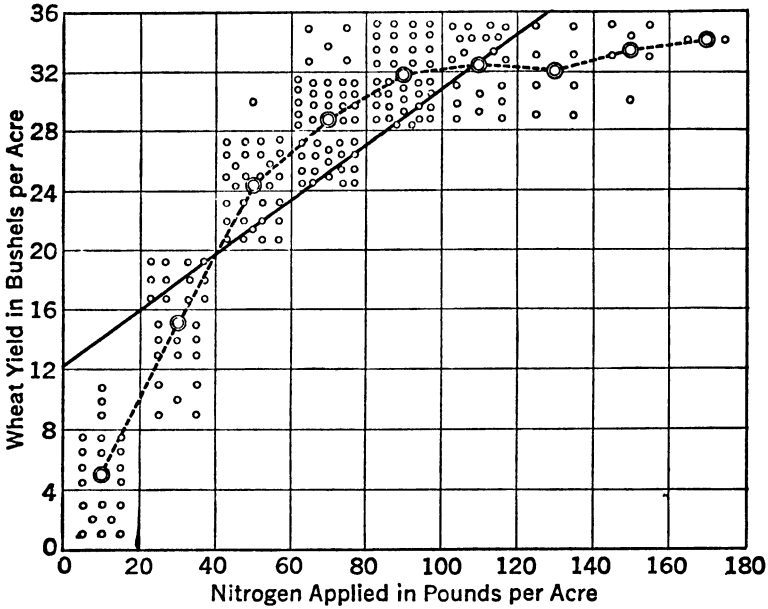


FIG. 83. — Scatter Diagram Showing the Relation between Wheat Yield and Nitrogen applied as Fertilizer, with Straight Line of Regression and Line joining the Means of the Columns

in calculating S_y, that is, by measuring and squaring the deviations of the individual points from the line of regression. In the present case, however, the line describing the relationship passes through the mean of each column, hence these means may be used in place of the "normal" values as computed from an equation of regression. In computing σ_{ay}, therefore, the deviations of the individual

actual figures used have been arbitrarily chosen for the purpose of the present illustration, but Davenport's experiments have demonstrated the existence of a law similar to the one here assumed.

items from the means of the various columns are squared added, the mean determined and the square root extracted, just as in the computation of the standard deviation. Part of the procedure is illustrated in Table 105, using the data in the first column of Table 104. This column contains all items having X-values between 0 and 20. The mean Y-value of the 21 items falling in this column is 5.05; deviations are measured from this value.

TABLE 105

Computation of the Squares of the Deviations about the Mean of an Array

Class-interval (wheat yield in bu. per acre)	m	f	*Deviation from mean of column (5.05)* d	d^2	fd^2
8-11 9	10	3	4.95	24.5025	73.5075
4- 7 9	6	10	.95	.9025	9.0250
0- 3.9	2	8	− 3.05	9.3025	74.4200
Total					156.9525

The sum of the squared deviations is obtained for each of the other columns in a similar fashion. The standard deviation about the means of all the columns, σ_{ay}, is found to have a value of 2.420. The value of σ_y is 9.188.

Substituting the given values in the formula

$$\eta_{yx}^2 = 1 - \frac{\sigma_{ay}^2}{\sigma_y^2}$$

$$\eta_{yx}^2 = 1 - \frac{(2.42)^2}{(9.188)^2}$$

$$= 1 - .0694$$

$$= .9306$$

$$\eta_{yx} = .965.$$

This is the value of the correlation ratio, measuring the degree of scatter about a line running through the means of the columns. Its significance is discussed below.

The method of calculation employed in the preceding example may be materially shortened. Let σ_{my} represent

the standard deviation of the means of the various columns about the arithmetic mean of all the Y's. In computing this value the mean of each column is weighted by the number of items in that column. It may be shown[1] that

[1] The following proof is adapted from Yule.

Given a series with mean M made up of two component series with means M_1 and M_2. N, the total number of observations, is equal to $N_1 + N_2$, the sum of the observations in the two component series. What is the relation between σ, σ_1 and σ_2? If we let $M_1 - M = c_1$ then for S_1^2, the mean-square deviation of the observations in the first of the two component series, measured from M as origin, we have

$$S_1^2 = \sigma_1^2 + c_1^2.$$

Similarly $$S_2^2 = \sigma_2^2 + c_2^2.$$

But $N_1S_1^2$ is equal to the sum of the squares of the deviations, about M, of the items in the first of the component series, and $N_2S_2^2$ is equal to the sum of the squares of the deviations, about M, of the items in the second of the two component series. Therefore

$$\sigma^2 = \frac{N_1S_1^2 + N_2S_2^2}{N}$$

and $$N\sigma^2 = N_1S_1^2 + N_2S_2^2. \tag{1}$$

But $S_1^2 = \sigma_1^2 + c_1^2$ and $S_2^2 = \sigma_2^2 + c_2^2$

therefore $$N\sigma^2 = N_1(\sigma_1^2 + c_1^2) + N_2(\sigma_2^2 + c_2^2). \tag{2}$$

In the present case we have the major series with mean represented by M_y, and a number of component series (the items arranged by columns) with means represented by m_{y1}, etc. Let S_{ay} represent the standard deviation of any column of Y's about the mean of that column. Then we have a number of component series, with standard deviations S_{ay1}, etc., and with means differing from the mean of all the Y's by $M_y - m_{y1}$, etc. Substituting in equation (2), we have

$$N\sigma_y^2 = n_1[S_{ay_1}^2 + (M_y - m_{y_1})^2] + n_2[S_{ay_2}^2 + (M_y - m_{y_2})^2] + \ldots \tag{3}$$

$$N\sigma_y^2 = \Sigma n[S_{ay}^2 + (M_y - m_y)^2]. \tag{4}$$

But $$N\sigma_{ay}^2 = \Sigma(n \cdot S_{ay}^2)$$

for, in each column, $$S_{ay}^2 = \frac{\Sigma d^2}{n}$$

since d represents a deviation from the mean of that column. For all columns,

$$\sigma_{ay}^2 = \frac{\Sigma d^2}{N} = \frac{\Sigma(n \cdot S_{ay}^2)}{N}.$$

Substituting in equation (4)

$$N\sigma_y^2 = N\sigma_{ay}^2 + \Sigma n(M_y - m_y)^2. \tag{5}$$

By definition of the standard deviation of the means of the columns

$$\sigma_{my}^2 = \frac{\Sigma n(M_y - m_y)^2}{N}.$$

Therefore, from (5), $$\sigma_y^2 = \sigma_{ay}^2 + \sigma_{my}^2. \tag{6}$$

$$\sigma_{ay}^2 = \sigma_y^2 - \sigma_{my}^2.$$

Substituting for σ_{ay}^2 in the equation

$$\eta_{yx}^2 = 1 - \frac{\sigma_{ay}^2}{\sigma_y^2}$$

we secure

$$\eta_{yx}^2 = 1 - \left(\frac{\sigma_y^2 - \sigma_{my}^2}{\sigma_y^2}\right)$$

$$= \frac{\sigma_{my}^2}{\sigma_y^2}$$

$$\eta_{yx} = \frac{\sigma_{my}}{\sigma_y}.$$

Since σ_{my} may be much more easily determined than σ_{ay} the value of η is generally computed from this formula. The data of Table 104 may be used to exemplify the process. Calculations appear in Table 106.

TABLE 106

Illustrating the Computation of the Correlation Ratio

Type of array (X-value of items in column) (pounds)	*Mean value of Y-items in column (bushels)*	*Deviation from mean of all Y's (25.005)*	*Square of deviation*	*Frequency*	
	m_y	d	d^2	f	fd^2
10	5.05	− 19.955	398.202	21	8,362.242
30	15.12	− 9.885	97.713	25	2,442.825
50	24.40	− .605	.366	30	10.980
70	28.73	+ 3.725	13.876	44	610.544
90	31.73	+ 6.725	45.226	37	1,673.362
110	32.40	+ 7.395	54.686	20	1,093.720
130	32.00	+ 6.995	48.930	8	391.440
150	33.33	+ 8.325	69.306	6	415.836
170	34.00	+ 8.995	80.910	2	161.820
Total				193	15,162.769

$$\sigma_{my} = \sqrt{\frac{15{,}162.769}{193}}$$

$$= 8.864.$$

Substituting the given values in the formula

$$\eta_{yx} = \frac{\sigma_{my}}{\sigma_y}$$

we have

$$\eta_{yx} = \frac{8.864}{9.188}$$
$$= .965.$$

The process of computing the correlation ratio may be briefly summarized:

1. Arrange the items in the form of a correlation table.
2. Find the arithmetic mean of all the Y-items in each column (i.e., find the arithmetic mean of each Y-array of type X).
3. Compute the arithmetic mean of all the Y's.
4. Measure the deviation of the mean of each column from the mean of all the Y's. Square each of these deviations and multiply by the number of items in the given column. Get the sum of the squared deviations.
5. Divide this sum by the total number of items and extract the square root of the result. This gives the value of σ_{my}.
6. Compute σ_y.
7. Divide σ_{my} by σ_y. The quotient is η_{yx}.

The value of the correlation ratio of X on Y may be similarly computed, substituting the proper values in the formula

$$\eta_{xy} = \frac{\sigma_{mx}}{\sigma_x}.$$

The symbol σ_{mx} represents the standard deviation of the means of the various *rows* about the mean of all the X's. The value of the correlation ratio of X on Y depends upon the amount of scatter (horizontally) about the line joining the means of the rows. Its value will generally be different from that of the correlation ratio of Y on X. In the present case the value of η_{xy} is found to be .824. As the line of relationship approaches the linear form the two correlation ratios approach identity.

Like r, η can never exceed 1, this value being secured when there is no dispersion about the line joining the means of the columns (or rows). From the formula

$$\eta_{yx} = \frac{\sigma_{my}}{\sigma_y}$$

it is evident that the value of the correlation ratio is zero when σ_{my} is zero. This is the case when the mean of each column has the same value as the mean of all the Y's. Such a condition is found when an increase or decrease in the value of the X-variable brings no corresponding change in the value of the Y-variable. This means that in each column of the correlation table there is a distribution of cases similar to the general distribution of Y's. When this is true there is clearly no relation between the two variables.

The correlation ratio, it should be noted, never has a negative value. It is possible to determine by inspection of the correlation table, however, whether the relation between two variables is direct, or inverse, or a varying one.

The coefficient of correlation has one distinct advantage, as compared with the correlation ratio, in that when its value and the values of the two standard deviations are known the equations to the lines of regression may be readily determined. This is not true of η. To get a quantitative expression for the "law" of relationship between two variables, when η has been computed, an additional calculation for the purpose of fitting a curve to the means of the arrays would be necessary.

CORRECTION OF THE CORRELATION RATIO

The use of η is only possible when the data are numerous, and can be arranged in the form of a correlation table. If a limited number of items should be so arranged, and it chanced that there was but one item in each column, the two measures σ_{my} and σ_y would be identical and η would

necessarily have a value of 1. Computed from a very small number of cases and employing a large number of classes, the correlation ratio would be meaningless.

The raw correlation ratio may be corrected by the method employed on a preceding page for the index of correlation, with m set equal to the number of groups (i.e., to the number of columns, for η_{yx}; to the number of rows for η_{xy}). Thus, if $\bar{\eta}$ be the corrected value, we have

$$\bar{\eta}^2 = 1 - \left\{(1 - \eta^2)\left(\frac{N-1}{N-m}\right)\right\}.$$

In the present instance

$$\bar{\eta}^2 = 1 - \left\{(1 - .9306)\left(\frac{193-1}{193-9}\right)\right\}$$
$$= .9276$$
$$\bar{\eta} = .963.$$

The correction is very slight in the present case, but if N were small or m very large it would reduce the given value materially.

RELATION BETWEEN THE CORRELATION RATIO AND THE COEFFICIENT OF CORRELATION

When the relation between two variables is absolutely linear the line running through the means of the columns corresponds, of course, to the line upon which the coefficient of correlation is based. When this is the case η and r have the same value. As the relationship between the two variables departs from the linear form the values secured for η and r differ, η being always greater than r. This results from the fact that the scatter about a line joining the means of the columns will always be less than the scatter about a straight line fitted to these points, except when the straight line passes through every mean point. And the less the scatter about the line expressing the average relationship the greater the value of the measure of correlation. Thus for the alfalfa problem it was found

that r has a value of + .69, and that an index of correlation based upon a second degree parabola has a value of .80. The correlation ratio for the same material is .82. For the data of Table 104 the value of η_{yx} (uncorrected) was found to be .965; the value of r is + .793, the difference between the two being marked. The reason for the difference is found in Fig. 83, in which the straight line of regression of Y on X and the line joining the means of the columns are shown. The regression departs materially from linearity, and the scatter about the straight line of regression is much greater than the scatter about the line joining the means.

The relation between r and η affords a convenient test of linearity in a given instance, since the two values will be identical when the regression is strictly linear, and will differ the more widely the greater the departure from the linear form. The general test for linearity is

$$\zeta = \eta^2 - r^2.$$

Even in a case of linear regression it is probable that η and r will differ somewhat because of fluctuations due to chance alone. A material difference, as reflected in the magnitude ζ (zeta), indicates that a straight line does not describe the relationship in question and that r is not a suitable measure of correlation. In the example given above, in which η equals .965 and r equals .793, the measure ζ has a value of .302. (The uncorrected η is used in this test.) This is large enough to indicate that the regression is non-linear.

In later sections methods of testing for linearity are more fully discussed.

REFERENCES

Bowley, A. L., *Elements of Statistics*, Chaps. 6, 7.

Crum, W. L. and Patton, A. C., *An Introduction to the Methods of Economic Statistics*, Chap. 17.

Day, E. E., *Statistical Analysis*, Chap. 13.

Elderton, W. P., *Frequency Curves and Correlation*, Chap. 12.

Ezekiel, Mordecai, "Correlation," *Encyclopaedia of the Social Sciences*, Vol. 4.

Ezekiel, Mordecai, *Methods of Correlation Analysis*, Chaps. 6–8.

Kelley, Truman L., *Statistical Method*, Chap. 10.

Pearl, Raymond, *Medical Biometry and Statistics*, Chap. 14.

Pearson, Karl, "Mathematical Contributions to the Theory of Evolution (XIV). On the General Theory of Skew Correlation and Non-Linear Regression." *Draper's Company Research Memoirs, Biometric Series II.* 1905.

Pearson, Karl, "Notes on the History of Correlation." *Biometrika*, Vol. 13, 1920 (25–45).

Pearson, Karl, "On a Correction Needful in the Case of the Correlation Ratio." *Biometrika*, Vol. 8, 1911 (254–256).

Pearson, Karl, "On the Correction Necessary for the Correlation Ratio." *Biometrika*, Vol. 14, 1923 (412–417).

Rietz, H. L. (editor), *Handbook of Mathematical Statistics*, Chap. 8.

Tippett, L. H. C., *The Methods of Statistics*, Chap. 9.

Waugh, A. E., *Elements of Statistical Method*, Chap. 10.

Yule, G. U. and Kendall, M. G., *An Introduction to the Theory of Statistics*, Chap. 13.

CHAPTER XIII

ELEMENTARY PROBABILITIES AND THE NORMAL CURVE OF ERROR

Reference has been made in an earlier section to the family resemblance which is found among frequency distributions drawn from widely different fields. Attention was also drawn to a certain basic type, represented graphically by the symmetrical bell-shaped curve, which is called the "normal curve," or the "normal curve of error." In an earlier day this curve was looked upon as representing a fundamental law which described all distributions of quantitative data. From the modern standpoint this was quite an erroneous conception. The normal curve is viewed today as but one of a number of types of curves which may be used to describe frequency distributions. It is, however, by far the most important type. For many of the measurements used to describe distributions of observations (measurements such as the mean, the standard deviation, the coefficient of variation) are distributed in accordance with this normal law of error. The procedures employed in generalizing results obtained from the study of samples and, in particular, in determining the reliability of such generalizations, lean heavily upon this law. An understanding of the characteristics of the normal curve is essential to the statistician.

Elementary Theorems in Probability

We may approach this subject by a brief consideration of certain elementary principles of probability that enter into many forms of statistical work. A detailed explanation of the theory of probability would carry us beyond the

limits of the present volume. The treatment which follows is presented only as an introduction to the subject, designed to illustrate, by simple numerical examples, the relation between the principles of probability and the normal law of error.

In this argument we may use the following standard notation. If an event can occur in n ways, a of which are to be considered as successful and b as unsuccessful, the probability p of a successful outcome may be written

$$p = \frac{a}{n}$$

and the probability q of an unsuccessful outcome may be written

$$q = \frac{b}{n}.$$

Since the sum of the favorable and unfavorable outcomes is equal to the total number of events, we have

$$a + b = n.$$

Dividing by n,

$$\frac{a}{n} + \frac{b}{n} = 1$$

so that

$$p + q = 1$$

or *certainty*.

A probability, therefore, may be written as a ratio. The numerator of the fraction corresponding to this ratio represents the number of favorable (or unfavorable) outcomes, while the denominator represents the total number of possible outcomes.

EXAMPLES OF SIMPLE PROBABILITIES

If a coin be tossed, the turning up of a head being looked upon as a favorable outcome, we have, as the probability of a success,

$$p = \frac{1}{2}.$$

and of a failure,

$$q = \frac{1}{2}.$$

If we roll a die, regarding a six spot as a favorable outcome,

$$p = \frac{1}{6}$$

and

$$q = \frac{5}{6}.$$

If a card be drawn from a pack of 52 the chance of drawing the ace of spades is $\frac{1}{52}$, of failing in that endeavor, $\frac{51}{52}$.

THE ADDITION OF PROBABILITIES

What is the chance of securing *either* an ace of spades *or* a two of spades in a single draw from a pack of 52 cards? In such a case, where any one of several outcomes will be considered as favorable, the probability of a success is the sum of the separate probabilities. In this example

$$p = \frac{1}{52} + \frac{1}{52} = \frac{1}{26}.$$

The chance of drawing either a heart or a spade from a pack of playing cards is given by

$$p = \frac{13}{52} + \frac{13}{52} = \frac{1}{2}.$$

THE MULTIPLICATION OF PROBABILITIES

Two events are said to be independent when the outcome of one does not affect the outcome of the other. Thus the result of one throw of a die does not, presumably, affect the result of the next toss. The probability of a *compound event* (i.e., that two events, independent of one another, will both occur) is the *product* of the probabilities of the separate events. Thus the chance of securing an ace,

followed by a two spot, in two successive throws of a die, is given by

$$p = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}.$$

In computing the probability of a given outcome it is frequently necessary both to multiply and to add probabilities. For example, we wish to determine the chance of securing the total 5 from two dice thrown simultaneously. We may label the dice a and b to distinguish them. This total may be secured from any one of the four following combinations:

Die a	Die b
1	4
2	3
3	2
4	1

The chance of securing an ace with die a is $\frac{1}{6}$, of securing a 4 with die b is $\frac{1}{6}$. The chance of the two in combination is $\frac{1}{36}$. Similarly, the probability of each of the other three combinations is $\frac{1}{36}$. But any one of these four results will give a total of 5, and will be considered successful. Hence

$$p = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{9}.$$

We have in this example answered the question: What is the probability of securing exactly 5 in the toss of two dice? We might put the question: What is the chance of securing *at least* 5 in the toss of two dice? In this case a total of 5 or more will be considered a favorable outcome. Just as in the preceding example, we may work out the probability of securing each of the results which will be accepted as successful. The following summary indicates the probability of each of these totals:

Probability of throwing 12 with two dice $= \frac{1}{36}$

" " " 11 " " " $= \frac{2}{36}$

" " " 10 " " " $= \frac{3}{36}$

" " " 9 " " " $= \frac{4}{36}$

" " " 8 " " " $= \frac{5}{36}$

" " " 7 " " " $= \frac{6}{36}$

" " " 6 " " " $= \frac{5}{36}$

" " " 5 " " " $= \frac{4}{36}$

Sum of above probabilities $= \frac{30}{36}$

The chance of throwing at least 5 in the toss of two dice is, therefore, $\frac{30}{36}$ or $\frac{5}{6}$.

The Binomial Expansion and the Measurement of Probabilities

It is possible to express these facts in a generalized form. A simple illustration may be employed to exemplify the derivation of the general expression.

If two coins are tossed simultaneously there are four possible outcomes

a b a b a b a b
TT TH HT HH.

(The two coins are represented, respectively, by the letters *a* and *b*.) The chances of securing no heads, one head, and two heads are, respectively, $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$. If three coins (represented by the letters, *a*, *b*, and *c*) are tossed simultaneously, we have eight possible outcomes

a b c a b c a b c a b c a b c a b c a b c a b c
TTT TTH THH THT HTT HTH HHT HHH.

The chances of securing no heads, 1 head, 2 heads, and 3 heads are, respectively, $\frac{1}{8}$, $\frac{3}{8}$, $\frac{3}{8}$, $\frac{1}{8}$.

But these results may be derived without working out the separate probabilities in detail. We have employed p and q to represent, respectively, the probability of success and failure of a given event. If there are two independent events the compound probabilities are given by the expansion of the expression

$$(p + q)^2.$$

For the case in which p (e.g., the probability of throwing a head) $= q = \frac{1}{2}$, the probabilities of the various results are given by

$$\left(\frac{1}{2} + \frac{1}{2}\right)^2 = \frac{1}{4} + \frac{1}{2} + \frac{1}{4}.$$

These are the results secured in the first example cited in this section. If there are three independent events, with $p = q = \frac{1}{2}$, we have

$$\left(\frac{1}{2} + \frac{1}{2}\right)^3 = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8},$$

the probabilities secured in the second example.

If we wish to know not the separate probabilities but the probable frequencies of the various outcomes in a given number of trials, these may be computed from the expression

$$N(p + q)^n$$

where N represents the number of trials and n the number of independent events. Thus if there are 200 trials and there are two independent events, the probable frequencies are given by

$$200(p + q)^2 = 200(p^2 + 2pq + q^2).$$

With $p = q = \frac{1}{2}$ this gives us

$$200\left(\frac{1}{4}\right) + 200\left(\frac{1}{2}\right) + 200\left(\frac{1}{4}\right) = 50 + 100 + 50$$

which indicates the probable frequencies of 2 successes, 1 success, and no successes.

If there are three independent events, the probable frequencies in N trials are determined from the binomial expansion of

$$N(p + q)^3.$$

If N equals 200, we have

$$200(p^3 + 3p^2q + 3pq^2 + q^3).$$

If p equals $\frac{1}{2}$, we have

$$200\left(\frac{1}{8}\right) + 200\left(\frac{3}{8}\right) + 200\left(\frac{3}{8}\right) + 200\left(\frac{1}{8}\right).$$

These terms indicate, in order, the probable frequencies of 3 successes, 2 successes, 1 success, and no successes. The total frequencies secured by carrying through the process of multiplication will be equal to the number of trials, for all possible outcomes are covered by the expansion.

Thus, when we know in advance[1] the probabilities attaching to similar but independent events, we may determine the probable frequencies of any given number of successes or failures. This is true whether p and q be equal or unequal. It is necessary only that p and q remain constant. There is here a fact of great significance in the development of statistical theory.

A COMPARISON OF ACTUAL AND THEORETICAL FREQUENCIES IN THE REALM OF PURE CHANCE

Certain points of importance may be made clear by comparing some experimental results with the theoretical frequencies given by the binomial expansion. Twelve dice

[1] A distinction is generally drawn between *a priori* probabilities of the type described above, and *empirical* probabilities, knowledge of which is derived from observation or experience. As an example of the latter type we have, as the probability that a man aged 35 will live 10 years, the ratio $\frac{74,173}{81,822}$. This is based upon the American Experience Table of Mortality which shows that of 81,822 men living at the age of 35, there are 74,173 living ten years later.

were thrown a number of times. Each 4, 5, or 6 spot appearing was considered to be a success, while a 1, 2, or 3 spot was a failure. (In a typical throw we might have the following spots up: 3, 1, 5, 1, 2, 4, 4, 6, 3, 2, 3, 5. In this lot there are five successes, and the result is so tallied.) In a classical example recorded by W. F. R. Weldon[1] twelve dice were thrown in this way 4,096 times, a success being defined as above. The results are recorded in column (2) of Table 107, and the distribution is shown in Fig. 84. By computation we find the arithmetic mean and the standard deviation of this distribution to be, respectively, 6.139 and 1.712.

Let us compare with these results those which we might expect from the given conditions. Twelve dice were thrown each time, hence we are dealing with 12 independent events. There were 4,096 trials. Since either a 4, 5, or 6 is considered a success, $p = q = \frac{1}{2}$.

For the terms in the binomial expansion we have

$$(p+q)^n = p^n + np^{n-1}q + \frac{n(n-1)}{1\cdot 2}p^{n-2}q^2 + \frac{n(n-1)(n-2)}{1\cdot 2\cdot 3}p^{n-3}q^3 + \ldots + q^n.$$

In the present case we have

$$4{,}096\left(\frac{1}{2}+\frac{1}{2}\right)^{12}.$$

Expanding

$$4{,}096\left(\frac{1}{4{,}096} + \frac{12}{4{,}096} + \frac{66}{4{,}096} + \frac{220}{4{,}096} + \frac{495}{4{,}096} + \frac{792}{4{,}096} + \frac{924}{4{,}096} + \frac{792}{4{,}096} + \frac{495}{4{,}096} + \frac{220}{4{,}096} + \frac{66}{4{,}096} + \frac{12}{4{,}096} + \frac{1}{4{,}096}\right).$$

Completing the indicated multiplication we have the theoretical frequencies of the various possible successes in 4,096 throws of twelve dice. These are shown in column (3) of Table 107.

[1] Cited by F. Y. Edgeworth, *Encycl. Brit.*, 11th ed., Vol. XXII, 394.

TABLE 107

Comparison of Actual and Theoretical Frequencies in Dice-Rolling Experiment

(1) *Number of successes*	(2) *Observed frequencies*	(3) *Theoretical frequencies*
0	0	1
1	7	12
2	60	66
3	198	220
4	430	495
5	731	792
6	948	924
7	847	792
8	536	495
9	257	220
10	71	66
11	11	12
12	0	1
	4,096	4,096

The distribution of the theoretical frequencies is shown in Fig. 84, with that of the observed frequencies. The relationship of the two distributions is close.

When we have, as in this case, a knowledge of the probabilities involved, it is possible to determine the arithmetic mean and the standard deviation of the distribution of the theoretical frequencies. As a general expression for the mean number of successes, where the number of independent events and the probability of success are known, we have

$$M = np.$$

Applying the present values,

$$M = 12 \times \frac{1}{2} = 6.$$

The mean, as computed from the observed frequencies, is 6.139.

As a general expression for the standard deviation,[1] under the same conditions, we have

$$\sigma = \sqrt{npq}.$$

In the present case

$$\sigma = \sqrt{12 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{3}$$
$$= 1.732.$$

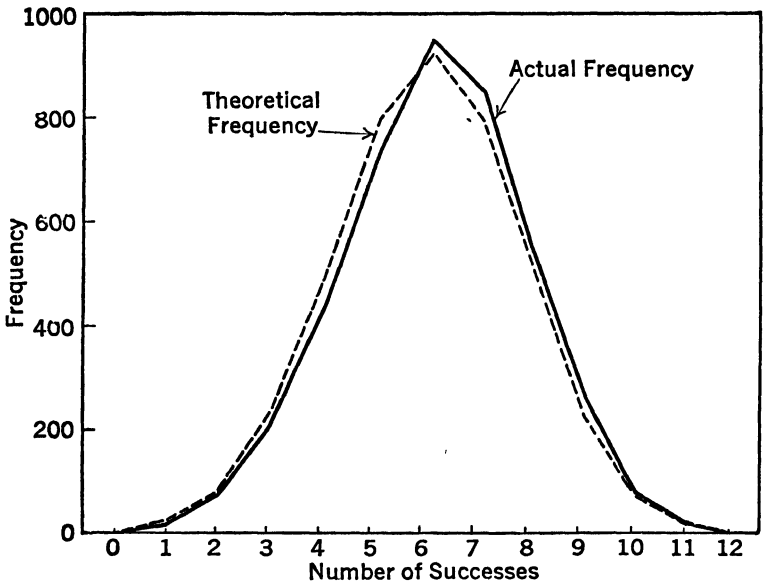


FIG. 84. — A Comparison of Actual and Theoretical Frequencies in a Dice-Rolling Experiment

The standard deviation, as computed from the actual frequencies, is 1.712.

When proportions, or relative frequencies, are dealt with, the standard deviation (σ') may be derived from the relation

$$\sigma' = \sqrt{\frac{pq}{n}}.$$

[1] This formula for the standard deviation of a binomial distribution is of central importance. The derivation of this formula, and that for the mean of a binomial distribution, are given in Appendix B.

The Normal Curve of Error

We may return to a consideration of the curve in Fig. 84 which represents the theoretical frequencies in the dice-throwing experiments. It is a perfectly symmetrical 12-sided polygon, the number of sides (excluding the base) corresponding to the number of independent events in the particular problem considered. With six events we should have a six-sided figure, with twenty events a twenty-sided figure, and so on. It is obvious that, as n increases, the number of sides to the polygon increasing correspondingly in number, the graph representing the expansion of the binomial $(p + q)^n$ approaches more and more closely a smooth curve. With n infinitely large a perfectly smooth curve would be secured. This is the normal curve of error which has been plotted in Fig. 85.

The equation to this curve is written in several forms, of which

$$y = y_0 e^{-\frac{x^2}{2\sigma^2}}$$

is one. In this equation y_0, the maximum ordinate, is a constant; e is a constant (the base of the Napierian logarithms) having a value of 2.71828; σ represents the standard deviation; and x is a given value of the dependent variable expressed as a deviation from the mean. The maximum ordinate may be derived from the relation

$$y_0 = \frac{N}{\sigma\sqrt{2\pi}}$$

hence the equation to the normal curve may be written

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

where π is the constant 3.14159.

This equation may be derived in several ways.[1] One

[1] Gauss' deduction of the error equation may be found in all standard works on the theory of least squares. Cf. references at end of Appendix A.

procedure which throws light on the physical conditions giving rise to the emergence of a normal distribution, starts from three basic assumptions.

1. The causal forces affecting individual events are numerous, and of approximately equal weight.
2. The causal forces affecting individual events are independent of one another.
3. The operation of the causal forces is such that deviations above the mean of the combined results are balanced as to magnitude and number by deviations below the mean.

A great part of the power which modern statistical technique possesses is derived from the detailed knowledge of the characteristics of the normal or Gaussian curve. From prepared tables showing the fractional parts of the total area under the curve lying between ordinates erected at stated distances from the maximum ordinate, theoretical frequencies may be determined much more readily than by the laborious method based upon the binomial expansion.

USE OF A TABLE OF AREAS UNDER THE NORMAL CURVE

The entire area under a frequency curve is taken to represent the total number of frequencies. Given information as to the proportion of the total area within a given segment, it would be easy to compute the frequencies represented by this segment, or to determine the probability that a given observation from the population represented by the curve would fall within the limits of this segment. Prepared tables of the probability integral, of which Table 108 is an example, serve just this purpose, with respect to the normal curve. (A more detailed table than that here given is needed for accurate computation. Appendix Table I will serve most purposes.[1])

[1] Tables of areas under the normal curve, as calculated by Dr. W. F. Sheppard, are available in many publications. Cf. *Tables for Statisticians and Biometricians*, edited by Karl Pearson, Biometric Laboratory, University College,

TABLE 108

Area of the Normal Curve in Terms of Abscissa

(Giving fractional parts of the total area between y_0 and ordinates erected at varying distances from y_0)

x/σ	a	x/σ	a
0.0	.00000	2.0	.47725
0.1	.03983	2.1	.48214
0.2	.07926	2.2	.48610
0.3	.11791	2.3	.48928
0.4	.15542	2.4	.49180
0.5	.19146	2.5	.49379
		2.5758	.49500
0.6	.22575	2.6	.49534
0.7	.25804	2 7	.49653
0.8	.28814	2.8	.49744
0.9	.31594	2.9	.49813
1.0	.34134	3.0	.49865
1.1	.36433	3.1	.49903
1 2	.38493	3 2	.49931
1.3	.40320	3 3	.49952
1.4	.41924	3.4	.49966
1.5	.43319	3.5	.49977
1.6	.44520	3.6	.49984
1.7	.45543	3.7	.49989
1.8	.46407	3.8	.49993
1.9	.47128	3.9	.49995
1.96	.47500	4.0	.49997

Since the normal curve is symmetrical about the maximum ordinate, the values given in Table 108 apply to observations on both sides of the mean. In using such a table, deviations from the mean are first expressed in units of the standard deviation. (The term *normal deviate* is applied to such a quantity, that is, to a deviation from the mean of a normal distribution expressed in units of the standard deviation of that distribution.) The proportion

London; *Tables of Applied Mathematics*, J. W. Glover, Ann Arbor, Michigan, George Wahr; *Manual of Problems and Tables in Statistics*, F. C. Mills and D. H. Davenport, New York, Henry Holt and Co.

of the total area lying between any two ordinates may then be readily determined. For example: What proportion of the cases in a normal distribution lies between the maximum ordinate and an ordinate erected at a distance from the mean equal to $+\sigma$? Reading down the x/σ column to 1.0, we find the value .34134 opposite it. This, in ratio form, is the proportion of cases falling within the limits indicated.

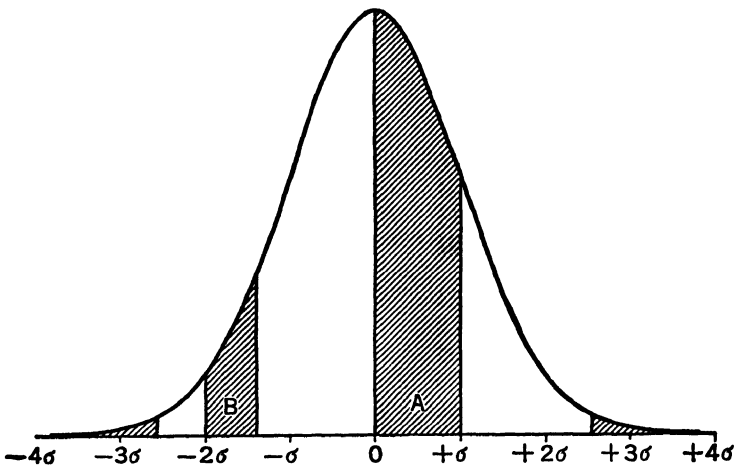


FIG. 85. — An Illustration of the Measurement of Areas under the Normal Curve

Expressing this ratio as a percentage, we have 34.134 per cent as the answer to our question.

Fig. 85 shows the relation of this area (the shaded area A) to the total area under the curve.

What proportion of the total number of cases in a normal frequency distribution will fall between an ordinate erected at a distance from the mean equal to -1.4σ and one erected at -2σ? From the table we find that 41.924 per cent of the total area will lie between y_0 and the ordinate at -1.4σ; 47.725 per cent will lie between y_0 and the ordinate at -2σ. The difference, 5.801 per cent, will fall between the ordinates at -1.4σ and at -2σ. This may be converted into actual frequencies by taking this propor-

tion of the total number of cases in the given distribution. The shaded segment B in Fig. 85 represents the area thus marked off.

For certain purposes we wish to know the proportion of the total number of cases deviating by a stated amount or more *in either direction* from the mean of a normal distribution. If we wish to know the proportion of all cases deviating from the mean by 1.96σ or more, we must add to the area between $+1.96\sigma$ and the upper limit of the curve the area between -1.96σ and the lower limit of the curve. Each of these areas equals .50000 − .47500, or .025. The percentage of cases deviating from the mean by $+1.96\sigma$ or more is 2.5; the percentage deviating by -1.96σ or more is 2.5. The percentage deviating above or below the mean by 1.96σ or more is 5.0. Similarly, it may be determined from the entries in Table 108 that just one per cent of all the cases in a normal distribution will deviate from the mean, positively or negatively, by 2.5758σ, or more. This "one per cent" area is represented by the sum of the shaded portions at the two tails of Fig. 85. The ordinates defining the inside limits of these segments are erected at $+2.5758\sigma$ and at -2.5758σ, while the outer limits are at infinity.

Special significance attaches to the two limits last mentioned, because of the uses made of them in interpreting errors of sampling. This topic is developed at a later point. Here we may note that the figures defining proportions of the total area under the normal curve falling in given areas may also be interpreted as probabilities. The probability that a given observation, made at random in a population distributed according to the normal law of error, will fall between the mean and a value one standard deviation above the mean is .34134; the probability that a given observation will deviate from the mean by 1.96σ or more is .05; the probability that a given observation will deviate from the mean by 2.5758σ or more is .01.

The method by which probabilities of occurrence may be determined from a table of areas under the normal curve, and by which the significance of a given normal deviate may be established, should be clearly understood. These methods enter in many ways into the work of a statistician.

The uses of the normal curve of error, and of the table of areas based upon the integration of this curve, are too varied to be enumerated at length here. A simple example may serve to introduce the subject.

An Economic Application

The statistical division of the American Telephone and Telegraph Company has made a study of the annual message use of four-party line residence message rate subscribers in Buffalo. The annual messages for each of 995 subscribers were tabulated and classified.[1] The results, together with certain computations, appear in Table 109.

The Moments of a Frequency Distribution

Some terms and symbols that have not been employed heretofore may be introduced at this point. We may write, using ν (nu) to define certain quantities of interest to us,

$\nu_1 = \frac{\Sigma f(x')}{N}$ = first moment of the distribution about the arbitrary origin.

$\nu_2 = \frac{\Sigma f(x')^2}{N}$ = second moment of the distribution about the arbitrary origin.

$\nu_3 = \frac{\Sigma f(x')^3}{N}$ = third moment of the distribution about the arbitrary origin.

$\nu_4 = \frac{\Sigma f(x')^4}{N}$ = fourth moment of the distribution about the arbitrary origin.

[1] "Introduction to Frequency Curves and Averages." *Statistical Bulletin, Statistical Methods Series, No. 1.* Issued by Chief Statistician, American Telephone and Telegraph Co.

TABLE 109

Annual Message Use of 995 Telephone Subscribers

(Illustrating the computation of the moments of a frequency distribution)

(1) *Interval of message use* *	(2) *Mid-point* m	(3) *Frequency* f	(4) *Deviation from arbitrary origin in class-interval units* x'	(5) fx'	(6) $f(x')^2$	(7) $f(x')^3$	(8) $f(x')^4$
0– 50	25	0	− 10	0	0	0	0
50– 100	75	1	− 9	− 9	81	− 729	6,561
100– 150	125	9	− 8	− 72	576	− 4,608	36,864
150– 200	175	19	− 7	− 133	931	− 6,517	45,619
200– 250	225	38	− 6	− 228	1,368	− 8,208	49,248
250– 300	275	50	− 5	− 250	1,250	− 6,250	31,250
300– 350	325	95	− 4	− 380	1,520	− 6,080	24,320
350– 400	375	85	− 3	− 255	765	− 2,295	6,885
400– 450	425	115	− 2	− 230	460	− 920	1,840
450– 500	475	132	− 1	− 132	132	− 132	132
500– 550	525	144	0	0	0	0	0
550– 600	575	116	1	116	116	116	116
600– 650	625	79	2	158	316	632	1,264
650– 700	675	54	3	162	486	1,458	4,374
700– 750	725	31	4	124	496	1,984	7,936
750– 800	775	11	5	55	275	1,375	6,875
800– 850	825	5	6	30	180	1,080	6,480
850– 900	875	6	7	42	294	2,058	14,406
900– 950	925	2	8	16	128	1,024	8,192
950–1,000	975	1	9	9	81	729	6,561
1,000–1,050	1,025	1	10	10	100	1,000	10,000
1,050–1,100	1,075	1	11	11	121	1,331	14,641
		995		− 956	9,676	− 22,952	283,564

"Moment" is a familiar mechanical term for the measure of a force with respect to its tendency to produce rotation. The strength of this tendency depends, obviously, upon the amount of the force and the distance of the point at which the force is exerted from the origin. The term is used in sta-

* As here classified an item having a value of 50 was put in the class having 50 as an upper limit. Items falling on other class limits were similarly disposed of.

tistics in a quite analogous sense, the class-frequencies being looked upon as the forces in question. The size of each class-frequency and the distance of each class midpoint from the origin are the factors of prime importance in this respect. The moments of a distribution about any origin may be computed by multiplying the frequency of each class by a given power of its distance, along the x-axis, from the origin, summing the resulting products and dividing by the number of cases. If the first moment is desired, the first power of the x-distance is employed; if the fourth moment, the fourth power of the x-distance, etc. The subscripts indicate the moments represented by the various symbols.

The most significant moments, for statistical purposes, are those which relate to the arithmetic mean as origin. Representing these moments by π (pi)[1] we have the following relationships:

First moment about the mean $= \pi_1 = 0.$
Second " " " " $= \pi_2 = \nu_2 - \nu_1^2.$
Third " " " " $= \pi_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3.$
Fourth " " " " $= \pi_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4.$

The computation of these moments from the data, as classified, involves the assumption that the items in each class can be treated as though they were concentrated at the midpoint of that class. It has been established that, under certain conditions, calculations made on this assumption are subject to a constant error. In particular, it has been shown that the values of the second and fourth moments are not the same, when computed from grouped data, as when computed from ungrouped data.

W. F. Sheppard[2] has worked out certain corrections for this bias. His corrections may be applied when two conditions prevail:

[1] In the equation to the normal curve π represents the familiar constant, 3.14159. As a symbol for a moment about the mean it relates, of course, to no such constant value.

[2] Cf. *Proceedings of the London Mathematical Society*, Vol. XXIX, 353–380.

(1) When the distribution relates to a continuous variable.

(2) When the frequency curve is characterized by "high contact," i.e., when the frequency curve tapers off gradually in both directions.

The symbol μ (mu) is employed to represent a corrected moment about the mean. The application of Sheppard's corrections gives us the following final formulation:

$$\mu_1 = 0$$
$$\mu_2 = \pi_2 - \frac{1}{12}$$
$$\mu_3 = \pi_3$$
$$\mu_4 = \pi_4 - \frac{1}{2}\pi_2 + \frac{7}{240}.$$

(In applying the corrections $\frac{1}{12}$ and $\frac{7}{240}$, the corresponding decimal values, .083333 and .029167, will generally be employed.) It is assumed in making these corrections that a class-interval unit has been employed in measuring deviations from the mean.

It may be noted in passing that the standard deviation is the square root of the second moment about the mean. For the uncorrected value,

$$\sigma = \sqrt{\pi_2}.$$

If Sheppard's corrections[1] are to be applied

$$\sigma = \sqrt{\mu_2}.$$

The calculation of the moments of the frequency distribution of telephone subscribers is shown on page 444. Sheppard's corrections are applied, since the curve is marked by reasonably high contact. It is a discontinuous distribution, but the unit (1) is so small in comparison with the range that it may be treated as continuous.

[1] It should be noted that these corrections, when appropriate, are applicable to the standard deviations entering into the calculation of the coefficient of correlation.

$$\nu_1 = \frac{-956}{995} = -.960804$$

$$\nu_2 = \frac{9{,}676}{995} = 9.724623$$

$$\nu_3 = \frac{-22{,}952}{995} = -23.067337$$

$$\nu_4 = \frac{283{,}564}{995} = 284.988945$$

$$\pi_1 = 0$$

$$\pi_2 = \nu_2 - \nu_1^2 = 9.724623 - .923144 = 8.801479$$

$$\pi_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3 = -23.067337 + 28.030370 - 1.773922 = 3.189111$$

$$\pi_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4$$
$$= 284.988945 - 88.652760 + 53.863384 - 2.556586$$
$$= 247.642983$$

$$\mu_1 = 0$$

$$\mu_2 = \pi_2 - \frac{1}{12} = 8.801479 - .083333 = 8.718146$$

$$\mu_3 = \pi_3 = 3.189111$$

$$\mu_4 = \pi_4 - \frac{1}{2}\pi_2 + \frac{7}{240} = 247.642983 - 4.400739 + .029167$$
$$= 243.271411$$

CRITERIA OF CURVE TYPE

Having these values, we may return to a consideration of the main problem, the utilization of our knowledge of the normal curve. There are certain criteria, represented by the letters β (beta) and κ (kappa), which enable us to determine readily whether a given distribution may be described by a curve of the normal type. These may be derived from the corrected moments of the given distribution.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{10.170429}{662.632015} = .01534853$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{243.271411}{76.006070} = 3.200683$$

$$\kappa_2 = \frac{\beta_1(\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)}$$

$$\kappa_2 = \frac{.01534853 \times 38.448470}{4(12.756686)(.355320)} = \frac{.5901275}{18.130823}$$

$$\kappa_2 = .032548$$

For the normal curve these criteria have the following values:

$$\beta_1 = 0$$
$$\beta_2 = 3$$
$$\kappa_2 = 0$$

We may conclude, tentatively, that the normal curve may be used to describe the given distribution.[1]

Fitting a Normal Curve; Use of a Table of Areas

The process of fitting a normal curve to a set of observations involves the computation of theoretical frequencies corresponding to the observed frequencies. This may be done from a table of areas under the normal curve (see Appendix Table I). Using such a table, in the manner indicated in the preceding section, the areas between the maximum ordinate and ordinates erected at the various class limits may be determined. By the simple process of subtraction the area within each class, and hence the theoretical frequencies, may then be computed. The procedure is illustrated in Table 110 on page 446, relating to the distribution of telephone subscribers.

The theoretical distributions derived from this fitting process may be compared with the observed frequencies, as given in Table 109. Or the comparison of the actual distribution and the fitted curve may be made graphically, as in Fig. 86. It is apparent by inspection that the normal curve gives a fairly good fit to the data, although there are several classes in which the differences are marked. A natural question arises as to the reason for the failure of the normal curve to fit at all points. There are two possible

[1] Account is later taken of the bearing of errors of sampling on this conclusion. See Chap. XVIII.

TABLE 110

Illustrating the Computation of Theoretical Frequencies from a Table of Areas

(1)	(2)	(3)	(4)	(5)	
Class limit	*Deviation from mean* $\frac{x}{\sigma}$	*Proportion of area between y_0 and ordinate at* $\frac{x}{\sigma}$	*Number of cases between y_0 and ordinate at* $\frac{x}{\sigma}$	*Theoretical frequencies, by classes*	
0	− 3.23	.4993810	496.88		
50	− 2.89	.4980738	495.58	0– 50	1.92*
100	− 2.55	.4946139	492.14	50– 100	3.44
150	− 2.22	.4867906	484.36	100– 150	7.78
200	− 1.88	.4699460	467.60	150– 200	16.76
250	− 1.54	.4382198	436.03	200– 250	31.57
300	− 1.20	.3849303	383.01	250– 300	53.02
350	− .86	.3051055	303.58	300– 350	79.43
400	− .52	.1984682	197.48	350– 400	106.10
450	− .18	.0714237	71.07	400– 450	126.41
500	+ .16	.0635595	63.24	450– 500	134.31
550	+ .495	.1896931	188.74	500– 550	125.50
600	+ .83	.2967306	295.25	550– 600	106.51
650	+ 1.17	.3789995	377.10	600– 650	81.85
700	+ 1.51	.4344783	432.31	650– 700	55.21
750	+ 1.85	.4678432	465.50	700– 750	33.19
800	+ 2.19	.4857379	483.31	750– 800	17.81
850	+ 2.53	.4942969	491.83	800– 850	8.52
900	+ 2.87	.4979476	495.46	850– 900	3.63
950	+ 3.20	.4993129	496.82	900– 950	1.36
1,000	+ 3.54	.4997999	497.30	950–1,000	.48
1,050	+ 3.88	.4999478	497.45	1,000–1,050	.15
1,100	+ 4.22	.4999878	497.49	greater than 1,050	.05
					995.00

answers to such a question. The failure to fit may be due merely to chance fluctuations such as are found in any sample. We may have an underlying law of distribution of residence subscribers, classified by message use, which

* The theoretical distribution shows .62 of a case below -3.23σ. To preserve formal consistency this amount has here been added to the theoretical frequency between 0 and 50.

accords perfectly with the normal law of error, but the particular sample selected may be marked by certain irregularities which would be ironed out if a very large number of cases were included. On the other hand, the differences may be due to the fundamental failure of such a distribution to accord with the normal law of error. Such a law may not describe the distribution of telephone calls, in which case the normal curve should not be employed.

At this stage we may note, without discussion, that the differences between theoretical and observed frequencies in

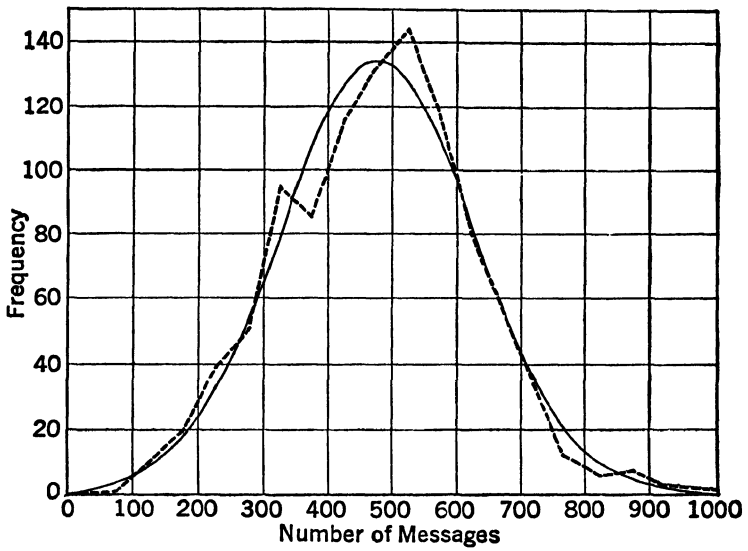


FIG. 86. — Illustrating the Fitting of a Normal Curve to Frequency Distribution of Telephone Subscribers, Classified according to Message Use

the present example are small enough to be attributed to chance fluctuations of sampling. The reasoning that supports this conclusion is presented in a later section (Chapter XVIII). The evidence is clear, however, that the discrepancies between the observed frequencies and those in the corresponding normal distribution are not excessively large. The observed facts are not inconsistent with the hypothesis

that residential telephone subscribers, classified according to frequency of telephone use, are distributed in accordance with the normal law of error.

This conclusion gives generality to the results of our study. We have a great deal of information concerning the attributes of distributions following the normal law of error, and once the identification of an actual distribution with this standard type has been effected we may draw upon this store of knowledge. In using the original frequency table we are limited to the classes there established. We may now go beyond this and determine how many cases may be expected within stated limits. We may compute the probability of a case falling between any two points on the x-scale, or above or below any given value. The observed results, standing alone, are restricted in their significance to the particular observations recorded, but the theoretical frequencies have no such limitations. They apply generally, to the entire population from which the sample was drawn. In so far as we are assured of the representative character of our sample we have a basis for inference that would be afforded by no amount of study of the particular distribution as a thing apart. This fact, that a knowledge of the theoretical frequencies permits *generalization* beyond the limits of direct observation, is perhaps the most important of the advantages derived from the identification of an actual distribution with an ideal type, such as the normal distribution.[1]

NOTE ON THE DESCRIPTION OF THE FREQUENCY DISTRIBUTION

With the aid of the criteria explained in this chapter it is possible to describe a frequency distribution more accurately than is possible with the measurements employed in the earlier chapters. A treat-

[1] As was stated, the normal curve is but one type of frequency curve, though one of basic importance. A comprehensive system of frequency curves is that associated with the name of Karl Pearson, who has derived equations to and has described in detail a number of standard types. An account of other fundamental types will be found in the books by Arne Fisher referred to at the end of this chapter.

ment of this subject is beyond the scope of the present book, but it seems advisable to indicate briefly the nature of these additional measures.

The value of β_2 serves as a measure of the degree of "flat-toppedness" found in a given curve. If $\beta_2 = 3$, as in the normal type, the curve is said to be *mesokurtic*. If $\beta_2 < 3$ the curve is *platykurtic*, or flatter than the normal type. If $\beta_2 > 3$, as in the example given above, the curve is *leptokurtic*, or more peaked than the normal.

A measure of skewness which is more accurate than those given early in the book may also be computed from these criteria. Karl Pearson has shown that the quantity

$$\chi = \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

serves as a measure of the degree of asymmetry of a given curve. Inserting the values of β_1 and β_2 given above we have, in the case of the distribution based on message use,

$$\chi = -.05558.$$

(χ is positive if the mean is greater than the median, negative if the mean is less than the median. In the present case the value of the mean is 476.96, that of the median is 482.39, hence the skewness is negative.)

Finally, the distance, d, between the mean and the mode may be determined from the relation

$$d = \chi \times \sigma.$$

In the distribution described above (relating to telephone use) σ, in original units, equals 147.65. Hence

$$d = -.05558 \times 147.65 = -8.21.$$

Since

$$Mo = M - d$$

we have

$$Mo = 476.96 + 8.21 = 485.17.$$

This gives a truer approximation to the modal value than any of the methods discussed in Chapter IV.

The methods exemplified in Table 109 and the accompanying text provide, therefore, a straightforward procedure for the measurement of the essential attributes of a frequency distribution. The mean and mode as measurements of central tendency, the

standard deviation as a measure of dispersion, χ as a measure of skewness, and $\beta_2 - 3$ as a measure of the degree of concentration of observations near the point of maximum frequency, may be computed directly from the first four moments of a distribution. These methods are available, of course, whether or not a study is to be carried to the point of determining and fitting a frequency curve of an appropriate ideal type. They are to be recommended for use in any systematic study of frequency distributions.

REFERENCES

Bowley, A. L., *Elements of Statistics.* Part II, Chap. 2.

Brunt, David, *The Combination of Observations.* Chap. 2.

Camp, B. H., *The Mathematical Part of Elementary Statistics.* Part I, Chaps. 2, 5. Part II, Chap. 1.

Carver, H. C., *Frequency Curves* (In *Handbook of Mathematical Statistics,* Rietz, H. L., ed., Chap.7).

Elderton, W. P., *Frequency Curves and Correlation.*

Fisher, Arne, *An Elementary Treatise on Frequency Curves. The Mathematical Theory of Probabilities.*

Forsyth, C. H., *An Introduction to the Mathematical Analysis of Statistics.* Chaps. 5, 8, 9.

Fry, Thornton C., *Probability and Its Engineering Uses.*

Jones, D. C., *A First Course in Statistics.* Chaps. 12, 18.

Kelley, T. L., *Statistical Method.* Chap. 5. (The Kelley-Wood Table of the Normal Probability Integral is given as an appendix.)

Mises, R. von, "Probability," *Encyclopaedia of the Social Sciences,* Vol. 12.

Nagel, Ernest, "The Meaning of Probability," *Journal of the American Statistical Association,* March, 1936.

Neyman, J., *Lectures and Conferences on Mathematical Statistics.* Graduate School, U. S. Department of Agriculture, Lecture 1.

Pearl, Raymond, *Medical Biometry and Statistics.* Chaps. 11, 12.

Pearson, Karl, *Tables for Statisticians and Biometricians.* (The *Introduction* to these tables will be found particularly useful.)

Richardson, C. H., *An Introduction to Statistical Analysis.* Chaps. 9, 10.

Sheppard, W. F., "On the Calculation of the Most Probable Values of Frequency Constants for Data Arranged According to Equi-distant Divisions of a Scale." *Proceedings of the London Mathematical Society,* Vol. 29, 1898.

Sheppard, W. F., "The Calculation of the Moments of a Frequency Distribution." *Biometrika,* Vol. 5.

Smith, J. G., *Elementary Statistics.* Part IV, Chaps. 16, 17.

Tippett, L. H. C., *The Methods of Statistics.* Chaps. 2, 4.

Waugh, A. E., *Elements of Statistical Method.* Chap. 6.

Wilks, S. S., *The Theory of Statistical Inference, 1936–1937.* Chap. 1.

Yule, G. U. and Kendall, M. G., *An Introduction to the Theory of Statistics.* Chap. 10.

CHAPTER XIV

STATISTICAL INDUCTION AND THE PROBLEM OF SAMPLING

The preceding pages have been devoted to an account of tools employed in statistical analysis. Examples illustrating the application of these tools to specific problems have been presented, but the emphasis throughout has been on technique. It is appropriate at this point that we stand off a distance, enlarging our perspective, and consider certain general problems relating to the application of these tools. What is their proper place in economic and business research? What are the assumptions involved in using them and what are their limitations? What are the end products of statistical analysis? How valid are the conclusions reached? What restrictions attach to such conclusions? We must give thought to such questions as these, if statistical methods are to be intelligently applied.

Statistical Description and Statistical Induction

In approaching this subject we must first make clear the distinction between *statistical description* and *statistical induction.* By employing the methods of statistics it is possible, as we have seen, to describe succinctly a mass of quantitative data. Hundreds or thousands of individual cases may be classified, and a frequency distribution formed. The essence of this distribution may be boiled down to perhaps four measures — of central tendency, variation, skewness, and kurtosis. A tremendous gain has been realized in thus replacing the multiplicity of individual cases by a limited number of measures that define the characteristics of the group as a whole. The possession of such tools makes

it possible for our limited powers of perception to grasp the significance of facts in the mass. Again, the methods of statistics enable us to describe relations between variable quantities. By securing the equation to an appropriate curve fitted to the data by mathematical methods, we may determine how much, on the average, one quantity changes in value as a related factor varies. This may be supplemented by a measure of the scatter or dispersion about the fitted curve, and by a measure, in abstract terms, of the degree of correlation between the dependent and the independent variables.

In so far as the results are confined to the cases actually studied, these various statistical measurements are merely devices for describing certain features of a distribution, or certain relationships. Within these limits the measures may be used with perfect confidence, as accurate descriptions of the given characteristics. But when we seek to extend these results, to generalize the conclusions, to apply them to cases not included in the original study, a quite new set of problems is faced.

The logical process by which one arrives at generalizations from a study of particular cases is termed *induction,* as opposed to *deduction,* which involves the drawing of specialized conclusions from general propositions. By *statistical induction* or *statistical inference* is meant the generalization of statistical results, the application to a *population* of measurements derived from a *sample.* We are employing this procedure constantly in practical statistical work, though not always with a full realization of the assumptions inherent in that process and of the limitations attaching to it.

The Nature of Statistical Induction

The problem at issue in considering the validity of statistical induction may be put in the following form: A statistical measurement — an average, a frequency ratio,

a coefficient of correlation — has been derived from the study of sample data drawn from a given population. (The term "population" refers to a complete universe of things or phenomena having stated characteristics in common.) May we assume that, if additional samples were taken from the same population, the corresponding measurements would have the same values? If not, may we determine the approximate limits to the fluctuations to be expected in these measures, as derived from successive samples? Here, obviously, is a problem of supreme importance. Karl Pearson has called it "the fundamental problem of practical statistics." If we cannot be assured of a certain degree of stability in the results secured from successive samples it would be quite invalid to generalize from the examination of a limited number of cases. No weight would attach to any study except one covering the entire universe of things or phenomena composing the given population. Yet such all-inclusive studies of economic phenomena are practically impossible. Index numbers of prices, of wages, of living costs, equations describing the relation between the production and prices of given commodities, coefficients of correlation between temperature and crop yield — all must of necessity be based on the study of samples. The problem of statistical inference, in the words of Oskar Anderson, is that of so utilizing the samples as to arrive at the best possible approximation to the characteristics of the universe.

We have noted that statistical inference is a special form of a general process of reasoning, induction. Two points are to be emphasized concerning inductive reasoning. First, the conclusion of any induction holds only in terms of probabilities. For such a conclusion, by the very definition of an induction, applies to cases not included in the observations. As opposed to deductive reasoning, in which the conclusion is implicit in the premises, induction yields a conclusion going beyond the premises. When all

the cases to be covered by the conclusions are included in the observations, the conclusion ceases to be an induction and becomes a descriptive statement. Accordingly, although induction is a highly fruitful means of adding to human knowledge, it is always hazardous. A leap in the dark is always involved, when we apply conclusions to cases not yet observed.

The justification for this leap in the dark, and this is the second point we wish to stress, is found in an assumption that there is a "limitation to the amount of independent variety" found in nature. While there is variation in nature, the degree of such variation is limited; there is some uniformity in all natural processes. When we are dealing with quantitative data this uniformity in nature is found in the stability of large numbers, as exemplified by the curious regularities in such phenomena as birth rates or death rates. Nature, in other words, is not marked by utter chaos; principles of regularity, order and stability appear in all natural processes, and these principles are strongly evident when we deal with masses of quantitative data. Therefore, when we generalize such a measure as an index number of wholesale prices, we do so on some such assumption as this: It is reasonable to suppose that, in the larger population to which this result is to be applied, there exists a uniformity with respect to the characteristic or relation we have measured. As a result of this uniformity we should expect statistical measurements derived from successive samples drawn from this population to fluctuate within definite limits.

It is evident that in making this assumption, in saying "It is reasonable to suppose . . . ," we are introducing an hypothesis which is incapable of complete verification by purely statistical methods. There is, thus, in every statistical induction, an *a priori* element. The statistical conclusion can never stand completely on its own feet. It must be endorsed by reason and judgment if it is to

carry conviction. If a high positive coefficient of correlation were secured from the study of a sample relating to banana importations and sales of new life insurance, this would not furnish convincing evidence of a causal relation, or a relation of contingency, between these two variables. There is no reasonable basis for assuming that, in the larger universe of phenomena from which the sample was drawn, there would be uniformity with respect to this relationship.

Statistical inference differs from the general process of induction in that a quantitative result is generalized. We seek to apply to a larger group — the population — the value of mean, standard deviation, or coefficient of correlation that has been computed from a sample. The measurement secured from the sample is an estimate of the corresponding measurement relating to the population. The direct task faced in such generalization is that of determining the limits within which these estimates would probably fluctuate, if based upon a number of different samples drawn from the same population. A number defining these limits will serve as a measure of the reliability of the given results, when generalized to apply to the population.

We should make clear at this point the sense in which the term "population" is used. When we speak of a population we are referring to an aggregate, whether of persons, things, or measurements, having certain common characteristics, or generated by a given system of causes. The term may refer to a hypothetical population from which a given sample may or may not have been drawn, or to a parent population of which a given sample is assumed to be representative. It may be a population of prices, or a population of cephalic indices; the term is not restricted to a population of persons. R. A. Fisher speaks of a "population of possibilities," referring to the possible results of an experiment many times repeated. Of high importance in statistics are populations of statistical measurements —

means, coefficients of variation, standard deviations, etc. It is proper to note that the populations to which most statistical results apply are infinite in size. Statistical generalizations relate to hypothetical universes containing infinite numbers of units. We assume a sample to be drawn, not from the finite population that might be covered by actual enumeration, but from the infinite population, or universe, that would be generated if the forces or system of causes that brought this sample into being were to operate without limit. (Statisticians have given some attention to special techniques, appropriate for dealing with a finite universe, but problems with which we do not here deal are faced in such applications.)

The principle of the *uniformity of nature* is assumed, of course, to apply to the universes from which our samples are drawn, if these samples are to be made bases of inductive generalizations. We must assume that these universes are stable, and that all their attributes are stable. An attribute of such a stable universe may not be exactly determined from the attribute of a single sample, but measurements defining the attributes of numerous samples drawn from the same universe will be distributed about the true value (i.e., that of the universe) in a systematic fashion. Each sample value is, of course, an estimate of the true value of the corresponding attribute of the population at large.[1] The precise determination of the characteristics of this distribution of estimates is essential to the determination of their reliability.

Having knowledge of this distribution we may determine the limits within which estimates derived from different samples of the same population may be expected to fluctuate. A measure of these limits will serve as a

[1] By convention, not yet generally adopted, but useful, the attribute of the population which is being estimated is termed a *parameter*, while the estimate of it is termed a *statistic*. Our certain knowledge is limited to statistics. We use this knowledge to the best of our ability to provide us with approximations to the true parameters which we can never know.

measure of the reliability of the given results, when generalized. Such a measure might be secured by the laborious process of studying a great many different samples, just as the dice were thrown 4,096 times in a preceding example. Thus we might desire to test the reliability of an average of weekly earnings of a certain class of workers. A first average might be secured from a sample composed of 250 individual records. This result might be tested by computing 499 additional averages, each based on 250 individual records. These 500 averages would not be identical in value, but if they were tabulated a frequency distribution closely approximating the normal type would be secured. From this distribution we might compute the mean of all the averages and the standard deviation of these averages. This standard deviation would serve as a measure of the variation found in the averages of weekly earnings, as computed from successive samples.

We have noted at an earlier point that a Gaussian or normal distribution is generated when three general conditions prevail. These are:

a multiplicity of forces affecting each observation
independence of the various forces affecting each observation
equality of the forces tending to generate values above and below the mean value.

The process of random sampling which would, presumably, be employed in securing the successive samples referred to in the preceding paragraph should satisfy the conditions giving rise to the normal distribution. There should be no special or unbalanced influences affecting particular samples. The differences between successive samples should be such as arise from a combination of forces, intermingled and not open to separate definition; that is, from "the mass of floating causes known as chance." If these conditions be met, and if the field of observation (i.e, the universe being

sampled) be homogeneous, the distribution of means computed from the successive samples would be normal.

This is a fact of high importance to statistical inference. In the realm of original observations, relating to persons, things, or events, normal distributions are the exception, rather than the rule. But the measurements which the statistician derives from successive samples, and which he employs in the inductive reasoning by which he generalizes his results, are far more frequently distributed in accordance with the Gaussian law. Much of the power of statistical instruments derives from this fact.

The statistical investigator is rarely in a position to build up a frequency distribution of constants derived from numerous samples. It is generally impossible to take 400 or 500 successive samples, in testing the reliability of a given measurement. As a practicable alternative a process of mathematical deduction is employed, in determining the characteristics of distributions of statistical measurements derived from random samples, drawn under stated conditions from given populations. An example of such mathematical deduction is provided by the derivation of the mean and standard deviation of a distribution generated under the following conditions:

p, the probability of a given event occurring, is known
q, the probability of the event not occurring, is known
n, the number of independent events in a single trial, is known.

Under these conditions, as was noted in the preceding chapter, $M = np$, and $\sigma = \sqrt{npq}$.[1] By a somewhat similar chain of reasoning, we may determine the characteristics of a distribution composed of arithmetic means of a number of samples of constant size drawn from a given population. The standard deviation of such a distribution is given by

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

[1] For proofs, see Appendix B

where σ_M is the required measure, σ is the standard deviation of the population from which the samples have been drawn, and N is the number of observations in each of the samples.[1]

The determination by deduction of the characteristics of distributions of statistical constants derived from samples is fundamental to the whole process of statistical inference It is not, of course, a task that needs to be done afresh in each statistical investigation. When the law of distribution of a given class of statistical measurements has been determined, statisticians may utilize the results in their various research fields, with due regard to all the conditions under which the given law holds This basic task has been performed for most of the statistical measurements currently employed. Earlier approximations have been refined in recent years for many classes of statistical measurements. The statistician today may draw upon a considerable body of tested and verified materials in determining the reliability of various kinds of statistical estimates. These materials exist in the form of shorthand expressions for the standard errors of different statistical constants, and in prepared tables for use when the distributions deviate materially from the type defined by the normal law of error.

Practical Problems of Sampling

The preceding discussion has dealt with one aspect of statistical induction. The argument has proceeded on the assumption that inferences concerning the attributes of a population would be based upon a sample thoroughly representative of the universe from which it was drawn. The securing of such a sample is a first condition of valid statistical induction. Practical problems of the first importance are faced in the actual field work of sampling. The procedures employed in such field work lie, in the main, beyond the scope of the present book, but it is desirable

[1] For proof see Appendix C.

that the general nature of sampling techniques be indicated. References given at the end of the chapter deal in greater detail with these procedures.

The task of securing an adequate sample calls, on the negative side, for an avoidance of bias in the individual observations and of preventable errors in schedules and tabulations. The term bias is applied to observational errors that are cumulative and non-compensating. Personal prejudices on the part of reporters, mental attitudes of which the subjects may be unconscious, or the mere physical conditions of observation may lead to persistent errors that distort samples. Errors in recording and tabulation are easier to detect. Training of enumerators and careful editing of schedules and tables will keep such errors to a minimum.

On the positive side sampling technique is directed toward the securing of a sample that is truly representative of the universe of inquiry. This is a major task, calling for a high degree of care and judgment in planning field operations conforming to the ultimate objectives of the study. A. L. Bowley has classified, under the four heads distinguished below, methods suitable for use in securing a representative sample.

The method of *random selection* is employed when the entire population to be sampled is treated as a whole, and members of the sample are so chosen as to be random members of that population. In this selection the individual choices must be independent of one another, and the chance of any member of the entire population being included in the sample must be the same as that of every other member. As regards the conditions of selection there should be present no element of preference or bias that would tend toward the inclusion or exclusion of certain members of the larger group. The general requirement here laid down should be interpreted, as J. M. Keynes has pointed out, to mean that *with respect to the purpose*

of the particular investigation the members of the sample should be random members of the population at large. Intelligent planning is needed in securing a purely random sample. The obvious procedure of picking the most readily available cases would by no means meet the condition of random selection. Certain important elements in the universe of facts to which the conclusions are to be applied may be excluded through the play of an unconscious bias unless careful attention is given to the selection of cases.

The population from which a given sample is to be selected is often not homogeneous, with reference to the purpose of particular investigation. Slum districts and wealthy districts may both have to be covered, in a study of social or economic conditions. Agricultural districts differing materially in fertility may be included in a farm survey. If, by a process of stratification, the universe of inquiry may be broken into sub-groups individually more homogeneous than the total population, the reliability of sampling results may be substantially increased. Within each sub-group random selection may be employed. This method is termed *stratified random selection.* The size of each group in the sample should be proportionate to the relative importance in the total population of the stratum represented by that group. Where homogeneous sub-groups are secured by the process of stratification, and where the differences between the sub-groups are pronounced, this method is distinctly superior to that of random selection among the undifferentiated members of the population at large.

In using the third method, that of *purposive selection,* the statistician seeks to secure a sample having the same characteristics as the universe of inquiry in respect of one or more "control" factors. If these controls are highly correlated with the quantities that are the objects of investigation, this method of selection gives obvious support to generalizations based on the study of the sample. As in

stratified selection, sub-groups are employed. These sub-groups are chosen not at random, but in such a way as to possess, in the aggregate, the same attributes (e.g., means, standard deviations) as the population at large, in respect of the control factors. Deliberate manipulation, often through a process of trial and error, is necessary to effect this agreement between the sample and the totality.

When this method is employed the statistician must, of course, have information concerning the "controls" for the total population. The application of the method is restricted to fields in which such knowledge is available. Census type inquiries on population, agriculture, and manufactures provide such basic knowledge. Promising work has been done in purposive selection in dealing with agricultural data.

The fourth method, that of *stratified purposive* selection, represents a combination of the use of stratification to secure homogeneous sub-groups and of deliberate selection through the use of controls. Where data are open to such stratification, and where necessary controls are available, the combined procedures may profitably be employed.

When a representative sample has been secured, when errors and bias have been avoided, we may still expect the attributes of the sample to differ from those of the total population. The effects of fluctuations of sampling will still be present, so long as the coverage of the sample falls short of the universe of inquiry. We may only estimate the attributes of the population; we still face the uncertainties that inhere in induction. It is possible, however, to define with considerable precision the probabilities involved in statistical induction when the differences between the attributes of the sample and those of the total population are due to fluctuations of "simple sampling," that is, to the scrambled mass of causes that constitutes chance. Under these conditions it is possible to assign in advance limits within which we may expect statistical measures

derived from different samples of the same population to fluctuate. This means that we may apply to the population at large statistical measures secured from the study of a sample, not with confidence in their perfect stability, but with fairly definite knowledge of the margin of error involved in thus extending our results. Where the necessary conditions are fulfilled statistical induction is a valid procedure.

Use of Measures of Reliability

Measurements defining the sampling errors to which given statistical constants are subject are put to various uses. It is in order now briefly to review the standard errors of different statistical measurements, and to illustrate their applications.

SAMPLING ERRORS: THE MEAN

For the standard error of an arithmetic mean we have

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

where the symbol σ in the numerator of the right-hand term refers to the standard deviation of the population from which the sample is drawn and N is the number of observations in the sample. Actually, of course, we do not know the standard deviation of the population, but we use as an approximation to it the standard deviation of the sample. The approximation is acceptable except when the number of observations in the sample is small, in which case special treatment is needed.[1]

Reference has been made above to the fact that a distribution of arithmetic means computed from random samples of a given population usually follows the normal law of error. This is true even though the distribution of the population from which the samples are drawn is not itself normal. Accordingly, we may interpret given

[1] See Chapter XVIII.

values of σ_M with reference to the probabilities associated with deviations in a normal distribution.

Table 34 in Chapter V shows the distribution in 1933 of 11,404 workers in open hearth steel furnaces, classified according to their average hourly earnings. The arithmetic mean of this distribution is 50.14 cents; the standard deviation, which we may here represent by s, is 18.685 cents. Accepting this standard deviation as an approximation to the standard deviation of the population from which this sample was drawn,[1] we have

$$\sigma_M = \frac{s}{\sqrt{N-1}} = \frac{18.685}{\sqrt{11{,}403}} = .175.$$

The true mean of the hourly earnings of wage workers in open hearth furnaces in 1933 is not known. The figure 50.14 cents is our best approximation to it. If we should draw many samples, each the size of the one we have here, we should have many mean values normally distributed and centering, we may assume, at the true value. The standard deviation of this normal distribution we estimate

[1] The formula for the standard error of the mean, when the σ of the population is known, is given by $\sigma_M = \frac{\sigma}{\sqrt{N}}$. When the standard deviation of the population is replaced by that of the sample (s), as an approximation to the desired quantity, the formula for σ_M may be written

$$\sigma_M = \frac{s}{\sqrt{N}}, \text{ or } \sigma_M = \frac{s}{\sqrt{N-1}}.$$

The first of these is appropriate if s has been derived from the relation $s = \sqrt{\frac{\Sigma d^2}{N-1}}$ (where d is the deviation of a single observation from the mean); the second is appropriate if s has been derived from the relation $s = \sqrt{\frac{\Sigma d^2}{N}}$. In other words, N should be reduced by 1 either in the derivation of s or in the derivation of σ_M. If σ_M is derived from the d's of the original data, the single operation is summed up in Bessel's formula

$$\sigma_M = \sqrt{\frac{\Sigma d^2}{N(N-1)}}.$$

(See Whittaker and Robinson, *Calculus of Observations*, London, Blackie & Son, 1924, 205–206.) The reason for the reduction of N is discussed in Chapters XV and XVIII, in dealing with "degrees of freedom."

as .175 cents. Knowledge of this standard deviation, or standard error, enables us to set limits within which it is highly likely that the true mean lies. Any statements we make about the true mean are to be interpreted with reference to this figure.

We might, for example, on the basis of these results, make the flat statement: The true mean of the population lies between 49.965 cents and 50.315 cents. (The first of these limits is the sample mean minus one standard error; the second is the sample mean plus one standard error.) We may not assert that this statement is certainly true. It may be true or false. But if we continue indefinitely to draw samples from the population in question, computing the mean of each and the standard error of that mean, and if we make a statement about each similar to that made above, 68 out of 100 such statements will be true. (The actual numerical limits set by the different statements will differ, of course.)

It is possible to vary the statement according to the degree of probability we wish to work with. Thus we might say: The true mean of the population lies between 49.80 cents and 50.48 cents.[1] Of an indefinitely large number of such statements, each based on the study of a sample similar to the one before us, we know that 95 out of 100 would be true. This is the kind of knowledge we have about generalizations based on results obtained from samples.

The essential facts concerning the mean of the present sample and its reliability may be summarized in the statement: The mean hourly earnings of wage workers in open hearth furnaces in 1933 was (in cents) $50.14 \pm .175$.[2]

[1] $49.80 = 50.14 - (1.96 \times .175)$
$50.48 = 50.14 + (1.96 \times .175)$

Ninety-five per cent of the area under a normal curve is included within $M \pm 1.96\sigma$.

[2] The measure of sampling reliability here given is the standard error. The traditional usage, now less commonly followed, has been to give the probable error, which is .6745 times the standard error. In the present example the

The standard error of a mean is frequently used, not merely as an abstract measure of sampling reliability, but as an instrument for testing a given hypothesis. Such an hypothesis usually involves an assumed parent population, and the test centers about the question whether a given sample could have been drawn from this parent population. Let us assume that, on rational grounds, we have set up the hypothesis that the mean duration of business cycles is five years. We have observations relating to 77 cycles occurring in various countries during stages of rapid industrialization.[1] These cycles are distributed, in respect of duration, as follows:

Duration of cycles, in years	*Number of cycles*
1	3
2	10
3	22
4	15
5	12
6	8
7	2
8	2
9	2
10	1
	77

The mean duration of these 77 cycles is 4.09 years, and the standard deviation of the distribution is 1.88 years. For the standard error of the mean we have

$$\sigma_M = \frac{1.88}{\sqrt{77 - 1}} = .216.$$

Are these results consistent with the hypothesis that our sample of 77 cycles is drawn from a parent population

probable error of the mean is .118 cents. It is well, in any case, to specify the exact measure of reliability being used.

[1] Cf. W. C. Mitchell, *Business Cycles, The Problem and Its Setting*, New York, National Bureau of Business Research, 1927, 412–416; F. C. Mills, "An Hypothesis Concerning the Duration of Business Cycles," *Journal of the American Statistical Association*, December, 1926, Vol. 21, 447–457.

(i.e., a universe of cycles generated under similar conditions) with a mean duration of five years?

If we use M to represent the mean of the sample data, M_h to represent the hypothetical mean of the universe, and T to denote the deviation of our sample mean from the hypothetical mean, expressed in units of the standard error of the mean, we may write

$$T = \frac{M - M_h}{\sigma_M} = \frac{4.09 - 5.00}{.216} = \frac{-.91}{.216} = -4.21.$$

The figure .216 is, according to our hypothesis, the standard deviation of a distribution of arithmetic averages the mean value of which is 5.00. If we were drawing from such a distribution, the mean of our present sample would represent a departure of 4.21 standard deviations from the general mean. What is the probability of such a departure occurring merely as a result of chance? Consulting a table showing areas under the normal curve, we find that the area on one side of the mean, lying at a distance of 4.21 standard deviations or more from the mean, constitutes 1/100,000 of the total area under the curve. In terms of probabilities, this means that there is only one chance in 100,000 that a member of the population represented by the normal curve will fall below the mean value by 4.21 standard deviations or more. This chance is so remote that we say the event in question could not occur. With reference to the present problem, we conclude that the results are not consistent with the hypothesis. We could not have secured the sample values in question had we been drawing from a universe of cycles with a mean duration of five years. The results fail to confirm the theory we have set up.

The probability cited (1/100,000) relates to a deviation on one side of the hypothetical value only. If we wish to define the probability of an observation departing from the hypothetical mean value 5.00 by 4.21 standard deviations or more, without reference to whether the departure be

above or below the hypothetical value, we must double the above probability. The chance of such a departure in one or the other direction is 2/100,000. Tests of hypotheses usually take this latter form. It is customary to ask whether a deviation of a stated magnitude could occur, and to measure the probabilities involved with reference to deviations in both directions.

In using tables of the normal probability integral in tests of this type we are generally concerned with the probability of occurrence of deviations as great as or greater than some stated value (in the above example, .91 years, or 4.21 standard deviations). This probability is represented by areas in the two tails of a normal curve (assuming that deviations either above *or* below the mean are in question). The inside limits of these segments are set by ordinates erected at distances from the mean equal to the deviation in question; the outside limits are at infinity. (See Fig. 85, in Chapter XIII, for a graphic representation of segments lying beyond stated limits.) The usual tables of the probability integral define the areas falling *within* limits set by ordinates at specific points. Our concern is with areas beyond these ordinates. Subtraction of the internal area from the whole area (unity) will, of course, give the area of the external portion defining the probability that is here desired.[1]

If we should be testing the hypothesis that the mean duration of business cycles is four years, we derive the value of T as follows:

$$T = \frac{4.09 - 4.00}{.216} = .42.$$

From the tabulated values of areas under the normal curve

[1] See W. Edwards Deming and Raymond T. Birge, "On the Statistical Theory of Errors," *Reviews of Modern Physics*, Vol. 6, July, 1934, 133ff., for a discussion of this probability, which they designate P_u, and tests based on it. (In their terminology, u is the difference between the mean of the sample and the mean of the assumed population.) This article includes a chart (134) for use in determining the significance of a given deviation.

we determine that approximately 67 per cent of all the observations in a normal distribution will deviate from the mean value by .42 standard deviations or more. We interpret this to mean that if our sample of 77 observations were drawn from a universe with a mean value of 4.00 years, the chances are 67 out of 100 that the mean of the sample would depart from the population mean by .09 years or more. (We have counted the combined probabilities of deviations above and below the population mean.) In other words, a deviation as great as the one we have experienced is highly probable. The results are not inconsistent with the hypothesis that the mean duration of business cycles is 4.00 years. They do not, be it noted, prove the hypothesis. All that we may say of statistical evidence, on the positive side, is that it is not inconsistent with a given hypothesis. Supporting statistical evidence strengthens our confidence in the hypothesis, of course. Its tenability must be determined on the basis of rational considerations as well as empirical evidence.

This last point deserves emphasis. "The significance of each test," say Deming and Birge,[1] "depends not only on the value of P (i.e., the measure of probability appropriate to the test) that is found, but also on how much is known *a priori* regarding the parent population." The above hypothesis of a four-year cycle has no particular rational basis (the figure was used here, of course, to exemplify a procedure). The fact that the observed results are not inconsistent with it is significant in a negative way, but does not establish the truth of the hypothesis. Low values of P, indicating that the facts are inconsistent with given hypotheses, are highly useful in leading us to reject tentative formulations of theory. Acceptable values of P, however, need the support of other knowledge (*a priori* and empirical) concerning the body of materials being studied and the regularities prevailing therein. Within the limit of acceptable

[1] *Loc. cit.*, 137.

values, indeed, we may accept one hypothesis, rather than another for which empirical tests yield a higher value of P, because the former is more consistent with the general body of existing knowledge concerning the field in question.

In the two tests we have applied, no difficulty was encountered in interpreting the probabilities bearing on the relation between the hypothetical mean and the observed facts. In the one case the odds were so small as to leave no doubt as to the lack of agreement; in the other case the difference was clearly insignificant. But many tests will lie on the borderline, and we must have some reasonable criterion as to the limit of significance. Odds of 1 out of 100 constitute one conventional standard. If a given difference between hypothetical and observed values would occur as a result of chance only 1 time out of 100, or less frequently, we may say that the difference is significant. This means that the results are not consistent with the hypothesis we have set up. If the discrepancy between theory and observation might occur more frequently than 1 time out of 100 solely because of the play of chance, we may say that the difference is not clearly significant. The results are not inconsistent with the hypothesis. The value of T (the difference between the hypothetical value and the observed mean, in units of the standard error of the mean) corresponding to a probability of 1/100 is 2.576. One hundredth part of the area under the normal curve lies at a distance from the mean, on the x-axis, of 2.576 standard deviations or more. Accordingly, tests of significance may be applied with direct reference to T, interpreted as a normal deviate (i.e., as a deviation from the mean of a normal distribution expressed in units of the standard deviation). A value for T of 2.576 or more indicates a significant difference, while a value of less than 2.576 indicates that the results are not inconsistent with the hypothesis in question.

There is, of course, nothing rigid about this particular

standard. Some statistical workers employ odds of 1 out of 20 as a limit, rather than 1 out of 100. With this standard we would accept as significant (i.e., not due to chance) a difference between hypothetical and observed values that would occur only 5 times out of 100, or less frequently, as a result of random fluctuations of sampling. The value of T corresponding to this standard is 1.96. The standards of significance actually employed by a research worker may well vary from problem to problem. The investigator uses the results of these tests of significance as aids in the interpretation of his results and in the development of a body of theory that is not inconsistent with the evidence provided by experience. In the interplay of deduction and induction that marks such a process, no single absolute standard for the rejection or acceptance of hypotheses would be appropriate.

The formula for the standard error of a mean, as given above, relates to a sample chosen by random selection. For a proportionately stratified sample the standard error of the mean, σ_{ms}, may be derived from the relation

$$\sigma_{ms}^2 = \sigma_0^2 - \frac{\sigma_m^2}{N}$$

where σ_0 is the standard error of the same mean as it would have been had the N observations been taken at random from the universe of inquiry, and σ_m is the standard deviation of the averages of the several strata about the average of the whole sample.[1] In computing σ_m the deviation of the mean of each stratum is weighted in proportion to the number of cases in that stratum. N is the total number of observations in the sample. It is clear from the formula that the standard error of the mean of the stratified sample is smaller than the standard error of a corresponding random sample.

[1] The above formula is from A. L. Bowley, *Elements of Statistics*, London, King, sixth ed., 1937.

SAMPLING ERRORS: MEDIAN AND QUARTILES

The median is subject to greater sampling fluctuations than is the mean. The degree of dispersion of median values derived from a number of samples of a stated size from a given population will be approximately 25 per cent greater than the dispersion of the arithmetic means of the same samples. More exactly, we have

$$\sigma_{Md} = 1.25331 \frac{s}{\sqrt{N-1}}.$$

Estimates of the quartiles, in turn, are less accurate than are estimates of the median. For these we have

$$\sigma_{Q1} = \sigma_{Q3} = 1.36263 \frac{s}{\sqrt{N-1}}.$$

SAMPLING ERRORS: STANDARD DEVIATION

In determining the magnitude of the sampling errors to which the standard deviation is subject we must distinguish between samples drawn from a normally distributed universe and those derived in the more general case, in which the nature of the distribution of the universe is unknown. If the distribution of the universe is normal we have, as the estimated standard error of σ,

$$\sigma_\sigma = \frac{s}{\sqrt{2N}}$$

(where $N-1$ has been used in the computation of s). Thus, for the universe of residential telephone subscribers represented by the distribution in Table 109, we have

$$\sigma_\sigma = \frac{147.7}{\sqrt{1{,}990}} = 3.31.$$

The more general formula for the standard error of the standard deviation involves the fourth as well as the second moment of the distribution:

$$\sigma_\sigma = \sqrt{\frac{u_4 - u_2^2}{4u_2 \cdot N}}.$$

For the distribution based on hourly earnings in open-hearth steel furnaces in 1933 the standard deviation was 18.685 cents (see Table 34). As the standard error of this measurement we have[1]

$$\sigma_\sigma = \sqrt{\frac{1{,}384.1183 - (13.9674)^2}{4 \times 13.9674 \times 11{,}404}} = .0432.$$

Since the moments here employed are in class-interval units, the derived measurement is also in those terms. In the original units we have

$$\sigma_\sigma = .0432 \times 5 \text{ cents} = .2160 \text{ cents.}$$

Many tests of significance involve the use of standard deviations and corresponding measurements of sampling reliability. These are discussed more fully in the chapter on the analysis of variance.

SAMPLING ERRORS: COEFFICIENT OF CORRELATION

A number of distinctive problems are faced in generalizing the results of correlation studies and in determining the significance of the measurements secured in such studies. Certain of these problems are discussed in the succeeding chapter, and Chapter XVIII deals with important limitations that are faced when the samples employed are small. At this point general methods of measuring the reliability of correlation measurements are presented, without certain of the qualifications that will be discussed later.

As a basic formula for the sampling error of the coefficient of correlation computed from N pairs of observations, we have

$$\sigma_r = \frac{1 - r^2}{\sqrt{N - 1}}$$

where r, in the numerator of the right-hand member, is the true coefficient of correlation in the population at large. Since we do not know the true r we must use the r of the

[1] Since Sheppard's corrections are not appropriate to this distribution, the uncorrected moments are used.

sample as an estimate of the required value. This formula may be taken to hold for distributions approaching the normal type, when the number of cases included in the sample is fairly large — say 50 or more. When the sample is small and, particularly, when we are dealing with a relatively high coefficient of correlation derived from a small sample, the standard error secured from the formula cited above may be faulty, and tests of significance based on it misleading. The reason for this and means of meeting the difficulty are discussed in Chapter XVIII.

In exemplifying the application of the usual test, we may employ results presented in Chapter X, on the relation between the discount rates of Federal Reserve banks and of commercial banks. The value of r is + .84, while N equals 1,800. Accordingly, we have

$$\sigma_r = \frac{1 - (.84)^2}{\sqrt{1{,}800 - 1}} = \frac{.2944}{42.40} = .007.$$

The standard error of r is frequently used, as are similar measurements relating to other statistical constants, to test hypotheses. We may put such a question as the following: Is the value of r secured from a given sample significant of a real relationship between the variables in question in the population from which the sample was drawn? Putting the question in form more appropriate for testing: Is the present value of r consistent with the hypothesis that there is no relationship between the variables in question in the population at large? R. A. Fisher terms such an hypothesis a "null hypothesis." The purpose of experiment, in his words, is to give the facts a chance of disproving the null hypothesis.

In a study of the movements of commodity prices, 1,202 measurements were secured on the timing of advances in the prices of individual commodities during periods of general business revival. Paired with each measurement was a similar observation on the timing of the decline in

the price of the given commodity during the succeeding period of general business recession.[1] We desire to know whether there is any relation between the sequence of price revival and the sequence of price recession. Is there a pattern in price movements during business cycles? Evidence of the existence of such a persistent pattern would lend support to the view that cycles represent true regularities in economic life.

These 1,202 pairs of observations yield a correlation coefficient of + .27. This does now show a pronounced degree of relationship. Our chief concern, however, is not with the magnitude of r. We wish to know whether the result is consistent with the hypothesis that the true correlation is zero. For the standard error or r we have

$$\sigma_r = \frac{1}{\sqrt{1{,}202 - 1}} = .029.$$

By hypothesis, the population value of r is zero, so the numerator of the fraction is 1.

If the true value of r were zero, and the standard error of r were .029, what would the probability be that, as a result of chance, we should secure a coefficient of + .27 from a given sample? Since this value represents a departure of more than 9 σ's from the hypothetical value of zero, the probability that the difference is due to chance is infinitely small. We conclude that the results are not consistent with the hypothesis that the sequence of price change during revival is unrelated to the sequence of decline in a succeeding recession. The null hypothesis is disproved.

Had the value of T $\left(\text{in this case } T = \frac{r - 0}{\sigma_r}\right)$ been less than 2.576 the conclusion would of course have been different. In such a case the discrepancy between the sample r and the hypothetical value of zero could be attributed to

[1] *The Behavior of Prices*, New York, National Bureau of Economic Research, 1927, 131.

sampling fluctuations. The result would not be inconsistent with the null hypothesis.

Having established that the results are not consistent with the hypothesis that the true value of r is zero, we may compute the standard error of r as actually derived. Assuming now that the sample is drawn from a parent population in which $r = +.27$, we have

$$\sigma_r = \frac{1 - (.27)^2}{\sqrt{1{,}202 - 1}} = .027.$$

SAMPLING ERRORS: INDEX OF CORRELATION

The standard error of the index of correlation may be approximated from the relation

$$\sigma_\rho = \frac{1 - \rho^2}{\sqrt{N - m}}.$$

In this formula m represents the number of constants in the equation of regression. In the example cited in Chapter XII, relating to alfalfa yield and depth of irrigation water, ρ is .80, N is 44, and m has a value of 3. We have, thus

$$\sigma_\rho = \frac{1 - (.80)^2}{\sqrt{44 - 3}} = .056.$$

The use and interpretation of this measure are analogous to those of σ_r. In the present instance the index of correlation is clearly significant.[1]

SAMPLING ERRORS: THE TEST FOR LINEARITY

As a test for linearity we have been given

$$\zeta = \eta^2 - r^2.$$

But we wish to know whether, in a given case, the difference

[1] See Ezekiel, M., *Methods of Correlation Analysis*, N. Y., John Wiley and Sons, 1930, 257–258, for a discussion of the sampling reliability of the index of correlation.

between η^2 and r^2 may be due merely to a chance fluctuation of sampling, or to a real departure of the underlying relationship from the linear form. As the standard error of ζ Blakeman has proposed

$$\sigma_\zeta = 2\sqrt{\frac{\zeta}{N}}\sqrt{(1-\eta^2)^2 - (1-r^2)^2 + 1}.$$

The use of this measure may be illustrated with reference to the problem relating to wheat yield which was considered in an earlier chapter. For the relation between wheat yield and amount of nitrogen used as fertilizer, we had

$$r = +.793$$
$$\eta = .965$$
$$N = 193.$$

(The uncorrected value of η should be used here.)

Therefore

$$\zeta = \eta^2 - r^2 = .302.$$

Inserting the given values in the formula for σ_ζ and solving, we have

$$\sigma_\zeta = .074.$$

With ζ having a value of .302, about 4.08 times its standard error, there can be no question as to the non-linearity of the relationship. The difference between η^2 and r^2 is one which could hardly be due to chance fluctuations of sampling.

The criterion $\eta^2 - r^2$ is not very satisfactory as a test of linearity, since the distribution of ζ does not follow the normal law. The same weakness attaches to the correlation ratio. As Fisher has demonstrated, the distribution of η does not tend to normality, even with large samples, unless the number of arrays is increased without limit. Accordingly, the standard error of η is of dubious utility. More efficient methods of testing for the existence of correlation, and for linearity, are discussed in Chapter XV.

SAMPLING ERRORS: COEFFICIENT OF RANK CORRELATION

The standard error of the coefficient of rank correlation has been given by "Student" as

$$\sigma_{\rho_r} = \frac{1}{\sqrt{N-1}}.$$

This formula holds on the assumption that the population value of the rank correlation is zero.[1] Under this condition the standard error may be taken to relate to a normal distribution, and interpreted in the familiar manner, when N is fairly large, say 45–50 or more. For small samples the distribution of ρ_r is not normal. In the example cited in Chapter X, dealing with the relation between the number of individual income tax returns and the number of passenger automobiles registered in 1934, by states, we had $\rho_r = .94$. Since there are 47 observations, the value of σ_{ρ_r} is given by

$$\sigma_{\rho_r} = \frac{1}{\sqrt{46}} = .147.$$

The sample is large enough to justify the assumption that the distribution of ρ_r would approximate the normal type. The coefficient of rank correlation is clearly significant, being more than six times its standard error.

SAMPLING ERRORS: COEFFICIENT OF REGRESSION

High importance frequently attaches to the coefficient of regression, in dealing with relationships among variable quantities. For the standard error of this measurement we have[2]

$$\sigma_b = \frac{s_y}{\sqrt{\Sigma x^2}}$$

where x is a given value of the independent variable,

[1] See Hotelling and Pabst, *loc. cit.*

[2] See R. A. Fisher, *Statistical Methods for Research Workers*, Edinburgh, Oliver and Boyd, sixth edition, 1936, 134–146.

expressed as a deviation from the mean of that variable, and s_y is the root mean square of the deviations of the actual values of y, the dependent variable, from the corresponding computed values. That is, s_y is a measure of the scatter about the line of regression.[1]

A test involving the use of σ_b may be applied to data relating to the average corn yield per acre in Kansas, by years, from 1890 to 1933 (see Table 128, Chapter XVI). These yields show a fairly consistent declining trend. A line of trend fitted to the figures for these 44 years is defined by the equation

$$Y = 22.05 - .1074X$$

where Y denotes corn yield per acre and X denotes time, in years, with origin at 1889. We wish to know whether the coefficient of regression (i.e., the slope of the line of trend) represents a significant departure from zero. The hypothesis we are testing is, then, that the true value of the coefficient of regression, in the population from which this sample is drawn, is zero — that there has been no significant decline in corn yield in Kansas over the period in question.[2]

For S_y we secure the value 6.70, for $\sqrt{\Sigma x^2}$ the value 84.23. Accordingly

$$\sigma_b = \frac{6.70}{84.23} = .0795.$$

We may denote by the symbol β the coefficient of regression

[1] $$s_y = \sqrt{\frac{\Sigma(y - y_c)^2}{N - 2}}$$

where y denotes a given value of the dependent variable and y_c denotes the corresponding value derived from the equation of regression. In the computation of s_y for this purpose N must be reduced by the number of constants in the equation of regression.

[2] The hypothetical population of which we assume our sample to be representative is the population that would be generated by the forces responsible for variation in Kansas corn yields from 1890 to 1933, if those forces, unchanged, were to act upon an infinite number of cases. The application of this concept, and of the whole probability calculus, to data ordered in time involves some logical difficulties, which are discussed at a later point.

assumed in our hypothesis (in this case zero). We wish to know whether the deviation of our actual b from this hypothetical β may be attributed to chance, or whether it is too great to be so explained. This deviation should be expressed in units of the standard error of b, in order that the probabilities underlying the normal distribution may be applied in our reasoning. Using T, as before, to denote the deviation in units of σ, we have

$$T = \frac{b - \beta}{\sigma_b} = \frac{-.1074 - 0}{.0795}$$
$$= -1.35.$$

The given value of b represents a departure of 1.35 standard deviations from the mean value of zero in our hypothetical population. As may readily be determined by reference to the table of the probability integral, such a deviation might easily occur, as a result of chance alone. The results then, are not inconsistent with the hypothesis. There is no clear evidence here of a significant decline in corn yield per acre in Kansas during the period covered.

SAMPLING ERRORS: DIFFERENCE BETWEEN MEANS

A problem of sampling that arises rather frequently is that of determining whether two samples could have been drawn from the same parent population. Obviously, there would be some difference between the means of two samples from the same universe, as there would be between standard deviations or coefficients of correlation secured from different sampling operations. We may illustrate the procedure employed in determining the significance of a difference between two arithmetic means.

Reference has been made above to a sample of 77 business cycles, occurring during stages of rapid industrialization. Their mean duration was 4.09 years; the standard deviation of the distribution was 1.88 years. The same investigation indicated that the mean duration of 51 business cycles

occurring in various economies during early stages of industrialization was 5.86 years, and that the standard deviation of these measurements was 2.41 years. There is an indication here that business cycles are accelerated, that their average length is shortened, when an economy is passing through a phase of rapid industrialization with corresponding impetus to technological change. In this case the null hypothesis against which we set our facts is that there is no difference, in respect of duration, between business cycles occurring in the two stages of industrialization named.

The difference between two means is a statistical measurement subject to a definite law of distribution. If a great many pairs of samples were drawn from a given population, the value D (i.e., $M_1 - M_2$) could be computed from the two means of each pair. A frequency distribution of the D's thus secured would follow the normal law. The magnitude of the standard deviation of this distribution would be a function of the sizes of the samples thus paired and of the standard deviations of these samples. We may approximate the standard deviation of this distribution of D's from the relation

$$\sigma_D = \sqrt{\frac{{\sigma_1}^2}{N_1 - 1} + \frac{{\sigma_2}^2}{N_2 - 1}}$$

or from

$$\sigma_D = \sqrt{{\sigma_{M_1}}^2 + {\sigma_{M_2}}^2}.$$

The measurement needed for testing the hypothesis now before us is computed from the relation

$$\sigma_D = \sqrt{\frac{(1.88)^2}{76} + \frac{(2.41)^2}{50}}$$
$$= \sqrt{.1627}$$
$$= .4034.$$

The value of D, the difference between the two means, is $5.86 - 4.09$, or 1.77. This value of D is to be judged

with reference to a hypothetical value of zero. Accordingly, for T (the discrepancy expressed as a normal deviate) we have

$$T = \frac{1.77 - 0}{.4034} = 4.39.$$

This discrepancy far exceeds the magnitude 2.576, corresponding to odds of 1 out of 100. If the true value of D were zero, a discrepancy as great as this or greater would occur as a result of chance about 1 time out of 100,000 trials. The results indicate that the difference between the two means is not due to chance. The facts are not consistent with the hypothesis that the two samples are drawn from the same population. There is a significant difference between the average durations of business cycles occurring in early stages of industrialization and in later stages of rapid industrial change.

SAMPLING ERRORS: DIFFERENCE BETWEEN PERCENTAGES

There are occasions when it is desirable to determine whether a difference between two proportions (or percentages) is significant. Using D_p to denote such a difference, we have

$$\sigma_{D_p}^2 = p_0 q_0 \left(\frac{1}{N_1} + \frac{1}{N_2}\right)$$

where p_0 is the weighted mean proportion, q_0 is $1 - p_0$, and N_1 and N_2 are the total numbers of cases in the two samples to which the proportions relate.[1] (In computing this value and applying the corresponding test it is necessary to divide percentages by 100, to reduce them to the form of proportions or ratios.)

A tabulation of American and foreign business cycles by Wesley C. Mitchell has indicated a relative preponderance of three-year cycles in American experience. Of 32

[1] See Hornell Hart, "The Reliability of a Percentage," *Journal of the American Statistical Association*, Vol. 21, March, 1926.

American cycles 10, or 31.2 per cent, lasted 3 years; of 134 cycles in other countries 20, or 14.9 per cent, lasted 3 years.[1] Is the difference between these two percentages great enough to justify the inference that the forces acting upon American business differ from those acting abroad, creating a significantly higher percentage of three-year cycles? The hypothesis that we test in this case is that the difference is not significant, that the groups of American and foreign business cycles are drawn from the same universe.

The two proportions, p_1 and p_2, with which we work are .312 and .149. The difference D_p between the two proportions is .312 − .149 or .163. For the weighted mean proportion we have

$$p_0 = \frac{N_1 p_1 + N_2 p_2}{N_1 + N_2}$$

$$= \frac{(32 \times .312) + (134 \times .149)}{32 + 134} = .1804.$$

$$q_0 = 1 - p_0 = .8196.$$

We compute the standard error of D_p from the relationship shown above

$$\sigma_{D_p}^2 = .1804 \times .8196 \left(\frac{1}{32} + \frac{1}{134}\right)$$

$$= .005724$$

$$\sigma_{D_p} = .0757.$$

Between the given value of D_p and the hypothetical value of zero we have the discrepancy (expressed as a normal deviate)

$$T = \frac{.163 - 0}{.0757}$$

$$= 2.15.$$

A discrepancy as great as this or greater might occur, as a result of chance, about 3 times out of 100. If our stand-

[1] See *Business Cycles, the Problem and Its Setting*, N. Y., National Bureau of Economic Research, 1927, 399–400.

ard of significance is 1 out of 100 we must conclude that the difference between the two percentages is not clearly significant. The result is not inconsistent with the hypothesis we set out to test — that American and foreign business cycles are drawn from the same universe, in respect of the proportion of three-year cycles occurring. It is proper to say, however, that we are dealing with border line results. If our standard of significance were 1 out of 20 we should consider the difference between American and foreign experience significant. Perhaps we should say that although the present evidence does not provide conclusive proof that the two samples come from different universes, there is indication of a difference between the forces affecting the relative frequency of three-year cycles in the United States and in foreign countries. Such results call for further research, in order that a more definite conclusion may be reached.

SAMPLING ERRORS AND SIGNIFICANT FIGURES

In deciding upon the number of figures to be recorded as significant, measures of sampling errors are, of course, pertinent. A useful general rule laid down by Truman L. Kelley follows: *In a final published constant, retain no figures beyond the position of the first significant figure in one third of the standard error; keep two more places in all computations.*[1] Its application may be illustrated with reference to the figures on hourly earnings of 11,404 steel workers in 1933. The mean, to four places, is 50.1360 cents. The standard error of the mean is .175 cents. One third of this is .0583. The first significant figure is in the column of hundredths. By the rule, therefore, the arithmetic mean should be given as 50.14 cents. Two more places, or four decimal places in all, should be retained in calculations.

[1] The rule here given is the Kelley suggestion as re-phrased by P. J. Rulon (*Science*, N. S. Vol. 84, No. 2,187, Nov. 27, 1936, 484). I have changed "one half the probable error" in Rulon's statement to "one third of the standard error."

Some Limitations to Measures of Sampling Errors

The importance of such measures of reliability as have been discussed above is, of course, great. With their aid we may give precision to our judgments concerning the margins of error involved in extending statistical results beyond the limits of actual observation. Yet limitations attach to them, and these must not be forgotten in a purely mechanical application of statistical tests.

Reference has been made to limitations relating to the size of samples. In the interpretation of most measures of sampling errors the assumption is made that statistical measurements secured from successive samples are distributed in accordance with the normal law of error. When the number of cases is large this is approximately true, even though the original data are not so distributed. But with a small number of cases in each sample this assumption may be quite invalid. The significance of given deviations (in terms of T) is therefore materially altered when we are dealing with results secured from small samples. Techniques have been developed, however, for defining sampling errors based on small samples. These are discussed at a later point (Chapter XVIII).

Moreover, the conventional standard errors we have discussed can be assumed to measure only errors arising from the fluctuations of simple sampling. If there is to be full conformity to the conditions of simple sampling, the probability of a given event occurring must be the same in all parts of the universe being sampled and for all time periods included, and the individual events (i.e., drawings or observations) must be completely independent of one another. The fact that customary error formulas are strictly applicable only when these conditions have been met injects elements of doubt into many statistical inductions in the field of economics. We cannot always be sure that the conditions of simple sampling are actually fulfilled.

They are rarely perfectly fulfilled in the handling of economic data. The standard errors derived above can give no indication of the possibility of fluctuations in successive samples due to causes other than those arising from simple sampling. Fluctuations due to bias, due to lack of representativeness in the sample, due to persistent errors of any sort, quite elude this method of determining probable stability. Although some degree of departure from the rigid conditions of perfect sampling does not deprive the measures of reliability of all value, the limitations noted must be the constant concern of the statistician.

The element of time adds one serious difficulty to the problem of statistical induction in the realm of economics, and in the social sciences generally. A universe that extends over time is subject to elements of change that are not present among data relating to a cross-section of time. Conditions of pig iron production, of banking, of foreign trade, of income distribution change from year to year, even from month to month. We may hardly assume that data relating to different time periods reflect the play of identical forces. When we deal with data from different periods we are, as Oskar Anderson has pointed out, drawing from different universes. The structural changes that occur in economic organization are manifestations of this state of never-ending transition. Accordingly the homogeneity of all populations extending over time is suspect. In particular are hazards faced when an induction extends to a time period not covered by the data of observation.

The fitting of trend lines, and the use of deviations from trend in statistical analysis, represent one effort to overcome difficulties arising out of temporal change. It is assumed that variations due to trend reflect the deep-seated changes that would introduce elements of heterogeneity into the particular universe of inquiry, and that deviations from trend may be made the bases of statistical inference. The effects of some temporal changes are doubtless removed by

this process. But the argument cannot justify the extension to a new time period of measures of sampling error based on the study of another period, unless it can be established that no essential change occurred in the conditions affecting the phenomena in question. The probable errors involved in such extension, without the validation noted, are not capable of definition. For this extension would involve generalizing about one universe from the study of another.

In the application of statistical methods proper choice of objectives, wise planning, and effective field work are of at least equal importance with skill in the use of statistical techniques. This is especially true as regards problems of sampling. Here chief emphasis falls on soundness and accuracy in the field work. The problems of field work are specialized and particular, arising out of specific problems and conditions. Appropriate special knowledge is needed for the selection and validation of the sample.

Much may be done to strengthen a statistical induction by making actual statistical tests of the homogeneity of the population and of the stability of sampling results. By the study of successive samples the representativeness of statistical measures may be determined; and by testing the subordinate elements of a given sample, when broken up into significant sub-groups, the inherent stability of a sample may be checked. The uniformity of nature in a given field is assumed in every induction. The induction is strengthened by every piece of evidence that supports the assumption.

REFERENCES

Anderson, Oskar N., "Statistical Method," *Encyclopaedia of the Social Sciences*, Vol. 14.

Bowley, A. L., *Elements of Statistics*, Part II, Chap. 4.

Bowley, A. L., "On the Precision Attained in Sampling," *Bulletin International Statistical Institute*, 1926.

Bowley, A. L., "The Application of Sampling to Economic and

Sociological Problems," *Journal of the American Statistical Association*, Sept. 1936.

Broad, C. D., "On the Relation between Induction and Probability," *Mind*, N. S. Vol. 27, 1918, and Vol. 29, 1920.

Burgess, Robert, *Introduction to the Mathematics of Statistics*, Chaps. 11–13.

Camp, B. H., *The Mathematical Part of Elementary Statistics*, Part II, Chap. 4.

Chaddock, R. E., *Principles and Methods of Statistics*, Chap. 11.

Cohen, Morris R., "The Statistical View of Nature," *Journal of the American Statistical Association*, June, 1936.

Davies, G. R. and Yoder, Dale, *Business Statistics*, Chap. 9.

Deming, W. Edwards and Birge, Raymond T., *On the Statistical Theory of Errors*, Graduate School, U. S. Department of Agriculture (reprint from *Reviews of Modern Physics*, July, 1937).

Editorial, "On the Probable Errors of Frequency Constants." *Biometrika*, Vol. 2 (273–281).

Elderton, W. P., *Frequency Curves and Correlation*, Chap. 10.

Fisher, Arne, *The Mathematical Theory of Probabilities.*

Jones, D. C., *A First Course in Statistics*, Chaps. 12–14.

Kelley, Truman L., *Statistical Method*, Chap. 5.

Keynes, J. M., *A Treatise on Probability.*

Mills, Frederick C., *On Measurement in Economics* (in *The Trend of Economics*, Tugwell, R. G. ed., 37–70).

Pearl, Raymond, *Medical Biometry and Statistics.* Chap. 10.

Pearson, Karl, "The Fundamental Problem of Practical Statistics," *Biometrika*, Vol. 13.

Pearson, Karl, *The Grammar of Science*, Chaps. 4, 5.

Richardson, C. H., *An Introduction to Statistical Analysis*, Chap. 11.

Rietz, H. L., *Random Sampling* (in *Handbook of Mathematical Statistics*, Rietz, H. L. ed., Chap. 5.)

Sarle, Charles F., "Reliability and Adequacy of Farm Price Data," *Bulletin* 1480, U. S. Department of Agriculture.

Smith, J. G., *Elementary Statistics*, Part IV, Chaps. 17–19.

Snedecor, G. W., *Statistical Methods*, Chap. 8.

Tippett, L. H. C., *The Methods of Statistics*, Chap. 3.

Wilson, E. B., "The Statistical Significance of Experimental Data," *Science*, Vol. 58, 1923.

Yule, G. U. and Kendall, M. G., *An Introduction to the Theory of Statistics*, Chaps. 18–21.

CHAPTER XV

THE ANALYSIS OF VARIANCE

The determination of degree of correlation between vari ables involves, essentially, the comparison of measurements of variability. Thus, in the familiar equation

$$r_{yx}^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$

we are comparing the dispersion about the fitted line of regression (S_y^2) with the dispersion about the mean of the y's (σ_y^2). Again, if we work with the relation

$$r_{yx}^2 = \frac{\sigma_{cy}^2}{\sigma_y^2}$$

we are comparing the dispersion of the computed values of y about the mean of the y's (σ_{cy}^2) with the dispersion of the original observations about the mean of the y's (σ_y^2). It is logical thus to compare measurements of variation, in applying correlation technique, for the purpose of the investigator is usually to test an hypothesis concerning the forces responsible for variation in the dependent variable. He is usually seeking an associated factor which may, on some rational basis, be assumed to influence the fluctuations of the variable he is treating as dependent. R. A. Fisher has developed a procedure to employ in the study of correlation which is based explicitly upon the analysis of variance. We deal in this chapter with certain applications of the flexible and powerful instrument Fisher has forged.

Comparison of Measures of Variability

We deal first with a simple comparison of two groups, in respect of variability. The prices of preferred and com-

mon stocks, as quoted on the New York exchanges, may be compared, to determine whether they differ significantly in variability Table 29, presented on a preceding page, showed the distribution of closing prices on July 25, 1936, of 66 preferred stocks, paying annual dividends of seven per cent. With this we may compare the distribution of a like number of common stocks selected at random from those for which prices were quoted on the New York Stock Exchange on July 25, 1936. The required values are given in Table 111.

TABLE 111

Comparison of Preferred and Common Stocks in Respect of Price Variation

	Degrees of freedom (n)	*Sum of squares of deviations from mean*	*Mean square deviation (variance)* σ^2	*Standard deviation* σ	*Common logarithm of standard deviation* $\log_{10} \sigma$	*Natural logarithm of standard deviation* $\log_e \sigma$
Common stocks	65	99,327.28	1,528.112	39.09	1.59207	3.66590
Preferred stocks (seven per cent)	65	30,812.20	474.034	21.77	1.33786	3.08056
					Difference =	0.58534

Each distribution includes 66 observations. (It is not essential to this comparison that the number of observations in the two distributions be equal.) In computing the mean square deviation we divide the sum of the squared deviations from the mean by n, the number of degrees of freedom, which is here equal to one less than the total number of observations in each distribution, that is, to $N-1$. (More is said below about the determination of number of degrees of freedom.) The standard deviation of the common stocks, 39.09, is materially greater than the corresponding figure, 21.77 for preferred stocks, but we cannot tell by inspection

whether the difference is significant, or whether it merely reflects a fluctuation of sampling. A precise test may be made by using the coefficient z as a measure of the difference in variability.

This coefficient is equal to the difference between the natural logarithms of the two standard deviations. That is

$$z = \log_e \sigma_1 - \log_e \sigma_2.$$

It is to be noted that natural logarithms are to be employed. Common logarithms on the base 10 may be shifted readily to natural logarithms on the base e (2.71828) by using the factor 2.3026 as a multiplier. From the entries in the last column of Table 111 we derive .58534 as the value of z.

If common and preferred stocks were alike, with respect to the dispersion of their prices, and if we had sufficiently large samples so that sampling fluctuations did not affect the measures of variance, the value of z would be zero. Is the value we have derived consistent with the hypothesis that the true value of z is zero? Could sampling fluctuations alone account for a deviation as great as .58534 from a true value of zero? If the derived value of z is too great to be attributed to sampling fluctuations, the hypothesis that common and preferred stocks are alike, with respect to the dispersion of their prices, is untenable.

To determine whether the derived value of z is consistent with the hypothesis that its true value is zero, we must know something about the distribution of values of z, if these were computed from many samples drawn under the same conditions. Fisher has shown that this distribution is normal, or effectively so, when the two distributions being compared both include a large number of observations. This is also true when the two distributions include only a moderate number of observations, but with n_1 and n_2 equal or nearly equal. The standard deviation of a dis-

tribution of z's secured under these conditions, or the standard error of z, is a function of the two n's. It may be derived from the relationship

$$\sigma_z = \sqrt{\frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where n_1 and n_2 are the number of degrees of freedom in the two distributions.

In the present example n_1 and n_2 are both equal to 65; the standard error of z is equal to the square root of the reciprocal of 65. We have

$$\sigma_z = \sqrt{.01538} = .124.$$

The test of the hypothesis that the true value of z is zero reduces, then, to the question whether a value of .58534 is likely to be drawn from a normally distributed population with a mean value of zero and a standard deviation of .124. A value of .58534 represents a deviation of 4.72 standard deviations from zero (i.e., $z/\sigma_z = 4.72$). A deviation as great as this occurs so seldom, in random sampling, that we may not accept the conclusion that the present value represents a chance deviation from zero. The result is not consistent with the hypothesis that the true value of z is zero. The dispersion of common stock prices is significantly greater than the dispersion of the prices of preferred stocks paying seven per cent dividends.

To exemplify a different condition, we may compare the dispersion of prices of preferred stocks paying six per cent and of preferred stocks paying seven per cent dividends. We have 64 quotations on the former, 66 quotations on the latter, both relating to closing prices on the New York Stock and Curb Exchanges on July 25, 1936. The figures are given in Table 112 on page 494.

In this comparison the value of z is .02890. The standard error of z (the square root of half the sum of the two reciprocals) is .12502. The coefficient z deviates from zero by

TABLE 112

Comparison of Six Per Cent and Seven Per Cent Preferred Stocks in Respect of Price Variation

	Degrees of freedom (n)	*Sum of squares of deviations from mean*	*Mean square deviation (variance)* σ^2	*Standard deviation* σ	*Natural logarithm of standard deviation* $\log_e \sigma$	$1/n$
Seven per cent preferred stocks	65	30,812.2	474.034	21.77	3.08056	.0153846
Six per cent preferred stocks	63	28,175.0	447.222	21.15	3.05166	.0158730
				Difference =	0.02890	Sum = .0312576

an amount equal to about one fourth of the standard error of z ($z/\sigma_z = .23$). This, of course, is a deviation that would occur very frequently in a normally distributed variate with mean value of zero. The result is, therefore, consistent with the hypothesis that the true value of z is zero. There is no significant difference between six per cent and seven per cent preferred stocks in respect of the dispersion of their quoted prices.

THE TESTING OF VARIABILITY BETWEEN CLASSES

The comparison of standard deviations provides a means of answering questions of another type. Measurements of changes in the average selling prices of products of manufacturing industries may be used to exemplify the procedure. If we classify manufacturing industries into those producing perishable, semi-durable, and durable goods, and compute an average of changes occurring between 1929 and 1933 in the selling prices of the products of each of these categories, we obtain the index numbers given in Table 113.

The average decline in prices was much less among durable manufactured goods than among goods of the other classes; semi-durable goods suffered the greatest loss. The range of variation among the three averages is considerable, but

TABLE 113

Measurements of Average Changes in Selling Prices, 1929–1933, in Three Groups of Manufacturing Industries

Class of industry	*No. of industries*	*Index of selling prices* 1929	1933
Producing perishable goods	34	100	69.81
Producing semi-durable goods	26	100	66.41
Producing durable goods	25	100	78.96
All industries	85	100	71.46

on the basis of the evidence here given we are not able to say whether the observed differences are due to chance, merely, or whether the prices of these several classes of goods were subject to the play of quite different forces, during the period here covered. An objective test is needed, before we may assume that the observed differences are significant.

For the application of such a test we need a measure of variation which is independent of the principle of classification here employed. How much might a series of price relatives for 1933, on the 1929 base, be expected to vary as a result of the play of chance? (By "chance" we here mean the mass of causes *unrelated to the factor of relative durability.*) A measure of the strength of such causes is provided by the variation *within* the three classes we have set up. The method used in measuring the variation within these classes is indicated in Table 114 on page 496.

It will be understood that the deviations which, in squared form, enter into the sums in the last column are the differences between individual items and the means of the classes in which those items fall. Thus the relative measuring the average selling price of products of the meat packing industry in 1933 was 44.90 on the 1929 base. This industry falls in the perishable goods group. The difference between 44.90 and 69.81 is 24.91. The square of this, or 620.5081, is one of the 34 items making up the

TABLE 114

Illustrating the Measurement of Variation within Classes

(1) *Class of industry*	(2) *No. of industries*	(3) *Mean of price relatives (1933 on 1929 base)*	(4) *Sum of squares of deviations of individual price relatives from class mean*
Producing perishable goods	34	69.81	6,464.0275
Producing semi-durable goods	26	66.41	3,375.1849
Producing durable goods	25	78.96	5,725.6916
All industries	85		15,564.9040

figure 6,464.0275, the entry for perishable goods in the last column of the preceding table. The sum of the entries in this last column, 15,564.9040, represents variation in price changes *within* the three classes. It is not influenced by factors of perishability or durability, since the total is affected only by variation *among* perishable goods, variation *among* semi-durable goods, and variation *among* durable goods.

Eighty-five items enter into this total. However, only 82 degrees of freedom are present. The 34 perishable goods possess 33 degrees of freedom to vary, the 26 semi-durable goods possess 25 degrees of freedom, and the 25 durable goods possess 24 degrees of freedom. For the standard deviation defining variability *within* classes we have, therefore

$$\sigma = \sqrt{\frac{15{,}564.9040}{82}} = 13.78.$$

This figure provides us with a yardstick, a measure of the degree of variation that is independent of the principle of classification employed in distinguishing perishable, semi-durable, and durable goods. This measures the variation due to the mass of floating causes known as chance.

With this standard we may compare the differences between the three class averages presented in Table 113. The magnitude of these differences may be defined by a single measurement, a standard deviation. In its computation the deviation of each class mean from the grand mean is measured, and the square of this deviation is multiplied by the number of items in the class in question. The procedure is illustrated in Table 115.

TABLE 115

Illustrating the Measurement of Variation between Classes

(1) *Class of industry*	(2) *No. of industries*	(3) *Mean of price relatives* (1929 = 100)	(4) *Deviation of class mean from mean of all observations*	(5) *Square of deviation of class mean from mean of all observations* $(4)^2$	(6) *Weighted squared deviation* (2) × (5)
Producing perishable goods	34	69.81	− 1.65	2.7225	92.5650
Producing semi-durable goods	26	66.41	− 5.05	25.5025	663.0650
Producing durable goods	25	78.96	+ 7.50	56.2500	1,406.2500
					2,161.8800

The sum of the entries in column (6), 2,161.8800, measures the total variation *between* classes. Although weights are used in getting this total, the differences relate to three separate averages, only, and but two degrees of freedom are represented in the total. As a measure of the degree of variation between the three broad categories we have set up, we have

$$\sigma = \sqrt{\frac{2{,}161.8800}{2}} = 32.88.$$

This we may take as a measure of the difference in degree of change in selling price from 1929 to 1933 that appears to be related to the relative durability of products.

The next step involves the formal testing of this figure against the standard provided by the measures of variation within classes. This test is applied in Table 116. Certain necessary calculations are also indicated.

TABLE 116

Comparison of Measures of Variation

Nature of variability	*Degrees of freedom* n	*Sum of squared deviations* [1]	*Mean square deviation* σ^2	*Standard deviation* σ	$Log_{10}\sigma$	$Log_e\sigma$
Between classes	2	2,161.8800	1,080.9400	32.88	1.51693	3.49288
Within classes	82	15,564.9040	189.8159	13.78	1.13925	2.62324
Total		17,726.7840		Difference $= z =$		0.86964

The test reduces, it is clear, to a comparison of two measures of variability. One, the standard of comparison, is the measure of variance *within* classes, a measure completely independent of the perishability or durability of the product. The other is a measure of variation *between* classes. Such variation might be due to the same general mass of causes responsible for variation within classes, or it might be due to special forces related to differences in the durability of the goods in question. If the former explanation is correct, the two measures of variation should be of the same order of magnitude, with due allowance

[1] The figure 17,726.7840, which is the sum of the squared deviations of individual observations from their respective class means and of the squared deviations of the several class means from the mean of all the observations, is equal to the sum of the squared deviations of all the individual observations from the mean of all the observations. In the table the total has been broken into two components, representing variability between classes and variability within classes.

for sampling fluctuations. If the second is the correct explanation the two measures of variation may differ appreciably in magnitude. The test, therefore, reduces to the question: Is the variation between classes significantly different from the variation within classes, account being taken of the degrees of freedom present in the two cases?

This question could be answered with reference to the standard error of z, provided the distribution of z be normal, or approximately normal. This is the case when the n's that measure the number of degrees of freedom are both large or when, though of moderate value, they are equal or nearly equal. This condition prevailed in the examples cited earlier. It is not met in the present instance, so we may not with accuracy employ the method of estimating and utilizing the standard error of z that was used in the earlier case. When the numbers of degrees of freedom are unequal and relatively small, as in this case, tests of significance may be most readily made with reference to a tabulation of values of z, prepared by R. A. Fisher. This tabulation gives, for various values of n_1 and n_2, values of z that would be exceeded 5 times out of 100, as a result of chance, if the true value of z were zero; it also gives one per cent values of z, i.e., values of z that would be exceeded 1 time out of 100, under conditions of random sampling, if the true value of z were zero. These two sets of values are reproduced in Appendix Tables VI and VII of this book, through the courtesy of Dr. Fisher and Oliver and Boyd, of Edinburgh, his publishers.[1]

In the present example the value of z, defining the degree of difference between the two measures of variation in

[1] Uses of the function z are discussed in R. A. Fisher's book, *Statistical Methods for Research Workers*, Edinburgh, Oliver and Boyd, sixth ed., 1936.

A table similar to Fisher's z-table, but relating to an alternative measure F, has been constructed by George W. Snedecor. F is derived directly from the variances (i.e., the values of σ^2) that are being compared; it is the ratio of the larger of the two variances to the smaller. For a table of values of F and a discussion of its uses see George W. Snedecor, *Statistical Methods*, Ames, Iowa, Collegiate Press, 1937.

Table 116, is .8696. In entering the z-table for the purpose of testing this measurement, the number of degrees of freedom corresponding to the larger of the two measures of variation compared is taken as n_1; n_2 is the number of degrees of freedom corresponding to the smaller of the two measures. This is a necessary procedure, with reference to the table as constructed. In the problem that now concerns us $n_1 = 2$, $n_2 = 82$.

For $n_1 = 2$ and $n_2 = 60$, the 1 per cent value of z is .8025; for $n_1 = 2$ and $n_2 = \infty$, the 1 per cent value of z is .7636. Interpolating, we obtain .7920 as the 1 per cent value of z for $n_1 = 2$, $n_2 = 82$.[1] If the true value of z were zero, we should expect a value as great as .7920, or greater, to occur as a result of chance only 1 time out of 100. The present value of z materially exceeds .7920; the probability of a value as great as this occurring as a result of chance, if the true value of z were zero, is less than 1 out of 100. The results of the test are not, therefore, consistent with the hypothesis that the true value of z is zero. The differences between the three class averages shown in Table 113 are too great to be attributed to chance. We may conclude that the price movements of perishable, semi-durable, and durable manufactured goods between 1929 and 1933 were significantly different.

[1] Interpolation in the z-table is based upon direct proportions among the reciprocals of the n's. In the above case

for $n_1 = 2$, $n_2 = 60$: the 1 per cent value of z = .8025	$1/n_2 = 1/60 = .0167$
for $n_1 = 2$, $n_2 = \infty$: the 1 per cent value of z = .7636	$1/n_2 = 1/\infty = .0000$
Δ = .0389	Δ = .0167

We must find the 1 per cent value of z corresponding to $n_1 = 2$, $n_2 = 82$. For $1/n_2$ we have

$$1/82 = .0122.$$

The difference between 1/82 and $1/\infty$, for which we must interpolate between the given values of z, is $.0122 - .0000 = .0122$. The required 1 per cent value of $z = .7636 + \left(\frac{.0122}{.0167} \times .0389\right) = .7920$.

The process of interpolation on the n_1 scale, if required, would be similar.

Variance Analysis in the Measurement of Relationship

The procedure employed in the comparison of measures of variability is applicable to the measurement of correlation. Indeed, using this technique it is possible to employ a systematic procedure that is of great value in revealing the character and degree of the relationships prevailing between variable quantities. This procedure is illustrated in the next section.

The method employed in applying to a typical correlation problem the method of analysis based on comparison of variances may be illustrated with reference to the data of alfalfa yield previously studied. These are presented in Table 117.

Table 117

Summary of Results Secured in Experiments with Alfalfa

(The measurements in the body of the table measure yields, in tons per acre, in 44 experiments)

	Inches of irrigation water applied								
	0	12	18	24	30	36	48	60	
	2.35	4.31	5.69	6.00	7.53	7.58	8 05	5.55	
	2.75	4.78	6.46	6.89	7.97	8 22	8.45	7.25	
	2.89	4.84	7.02	7.96	8.32	8 63	8.63	10 17	
	3 85	5.83	8.02	8.32	9.43	9.33	8.83	10.70	
	5.52	6.51		8.38	9 54	9.38	9 52		
Average	5.94	7.52	___	9.96	11.06	12.48	10 62	___	
yield	3.88	5.63	6.80	7.92	8.98	9.27	9.02	8.42	7.48

The average yield of alfalfa, in these 44 experiments, was 7.48 tons per acre. But there was rather wide variation among the results. The sum of the squares of the deviations of the 44 observations from the mean is 228.33. This sum sets our problem. We should like to find reasons for this variation.

TESTING FOR THE EXISTENCE OF CORRELATION

The observations are set up above in a form suited to the testing of one hypothesis concerning the factors affecting alfalfa yield. The data are arranged in eight arrays, classified according to the depth of irrigation water applied. This depth varied from 0 to 60 inches. Variations in yield appear to be associated with variations in amount of water applied. As a basis for our procedure we set up the hypothesis that there is no such association. To test this hypothesis, we may break the sum that measures the total variation of yields into two parts measuring, respectively, the variation within arrays and the variation between arrays.

To determine the total *variation within arrays*, the deviation of each observation from the mean of the array in which it falls is measured. The sum of the squares of these deviations, for all the arrays, is the desired total. Thus, in the first array of Table 117, the mean is 3.88 tons. The deviation of the first observation, 2.35, from this figure, is -1.53; its square is 2.3409. The deviation of the second observation, 2.75, is -1.13; its square is 1.2769. Determining in similar fashion the deviations of the four other observations in that array from the mean of the array, squaring these, and adding the six squared values, we have 11.5320 as the sum of the squares of the deviations in the first array. Performing similar calculations for the seven other arrays, and adding the eight sums thus secured, we have a figure of 76.39. This is the total variation within arrays. For convenience we may refer to this as component A of the total variation.

In determining the total *variation between arrays*, the deviations of the means of the various arrays from the mean of all the observations are measured and squared, and the weighted sum of these squares is secured. Weights are based upon the number of observations in the several arrays. Thus the mean of the first array, 3.88 deviates

from the mean of all the observations, 7.48, by 3.60; the square of this is 12.9600. Multiplying by six (the number of observations in the first class), we have 77.7600. Securing similar weighted figures for the seven other arrays, and adding, we have 151.94 as the total variation between arrays. This we may call component *B*.

In breaking up the total variation into two components[1] we have distinguished variations in yield that are definitely not related to differences in depth of irrigation water applied, from variations in yield that may or may not be related to irrigation differences. Within the first array, including six experiments on plots to which no irrigation water was applied, yields varied from 2.35 tons to 5.94 tons per acre. The total variation within this array (the sum of the squares of the deviations from the mean of the array) amounted to 11.5320. Since the irrigation factor was constant, this sum measures variation which is completely independent of changes in irrigation. This is true also of the figure 76.39, measuring total variation within all the eight arrays set up in Table 117. Differences in soils and innumerable minor factors combined to create variation within these arrays. The figure 76.39 measures the play of that host of undefined forces to which we give the name *chance*. The one specific factor which does not affect this figure is irrigation. We have measured the variation in such a way that irrigational differences do not enter.

Irrigational differences do enter definitely into the variation between arrays. Indeed, it may be the dominant factor in this variation, which is measured by the figure

[1] The sum of the two components is, of course, equal to the total.

Variation within arrays (Component *A*)	76.39
Variation between arrays (Component *B*)	151.94
Total variation	228.33.

For a demonstration of this relationship see note, pp. 418–9.

To ensure full consistency between components *A* and *B* and the total (and among the sub-divisions of *B* later defined), when these quantities are independently computed, it is necessary that all computations be carried to more decimal places than are customarily retained.

151.94. But of this we cannot be sure. For the means of the eight arrays differ among themselves not only because of differences in the amounts of irrigation water applied to the different plots. To yield differences due to the irrigation factor are added yield differences due to the innumerable other forces that influence alfalfa yield, the forces we lump together as chance. For chance factors affect the means of the various arrays, and so affect the variation between arrays, just as they affect the variation within arrays. As the experiment was designed, the influence of irrigational differences is present only in the variation between arrays, but the influence of "chance" is present in both the variation within arrays and the variation between arrays.

In this fact is found the key to our problem, and the instrument for testing our hypothesis. For, in so far as chance alone is operative, the variation between arrays would be expected to be of the same order of magnitude as the variation within arrays. The figures we have so far examined indicate that the variation between arrays is greater than the variation within arrays. But this may be a purely fortuitous result. The apparent increase of yield with increased irrigation may be entirely a chance phenomenon, similar to a run of heads in tossing a coin. This we must test. We must determine whether the forces responsible for variation between arrays are the same as the forces responsible for variation within arrays.

The hypothesis we shall test, and which may of course be disproved, is that the forces responsible for variation between arrays are the same as the forces responsible for variation within arrays; in other words, that there is no association between depth of irrigation water applied and alfalfa yield. The nature of the test to be applied has been indicated in the preceding sections. We shall compare two measures of variation, to determine whether they are of the same order of magnitude. But before this test is applied, account must be taken of the number of degrees of freedom pre-

vailing in each case. This concept calls for brief explanation.

If the data of alfalfa yield related to but one plot of land, in one year, there would be no variation. A single observation would coincide with the mean, and the standard deviation would be zero. With a second observation opportunity for variation arises. But we may think of it as a single opportunity. With but two observations there is but one degree of freedom to vary. With three observations, two opportunities to vary are given; there are two degrees of freedom. In problems of this sort the number of degrees of freedom is equal to $N - 1$. Our present example includes 44 observations; hence the total variation 228.33 represents the resultant of 43 degrees of freedom.

How are these 43 degrees divided between the two components, A and B? As regards variation within arrays, this may be readily determined by reference to Table 117. Variation within arrays, it will be recalled, was measured with reference to the means of the various arrays. In the first array, containing six observations, there exist five degrees of freedom to vary from the mean of that array. The same is true of the arrays relating to 12, 24, 30, 36, and 48 inches of irrigation water. In each of the arrays relating to 18 and 60 inches of water there are but four observations, with three degrees of freedom. The total of these degrees of freedom is 36. Variation between arrays was determined by measuring the deviations of the means of eight arrays from the general mean of the distribution. Since eight different values are involved, there are seven degrees of freedom. (The fact that weights were employed in securing the total variation between arrays does not affect the determination of degrees of freedom.) The 36 and the 7, combined, use up all the 43 degrees of freedom entering into the total variation.

Knowing these degrees of freedom we may now reduce the measures of variation within arrays and of variation

between arrays to comparable terms, and determine the significance of the difference between them. This is done in Table 118. This table and others following differ somewhat from those employed in similar comparisons in the opening sections of this chapter. In the earlier tables variability was measured in units of the standard deviation, and the function z was derived from the relationship

$$z = \log_e \sigma_1 - \log_e \sigma_2.$$

It is often more convenient to perform the necessary calculations in terms of the variance, that is, of σ^2, and to derive z from the relationship

$$z = \frac{\log_e \sigma_1^2 - \log_e \sigma_2^2}{2}.$$

The procedures lead to the same result, of course, since half the difference between the logarithms of the squared standard deviations is equal to the difference between the logarithms of the standard deviations, but the use of squared measurements eliminates one step in the calculation.

TABLE 118

A Test of the Existence of Correlation

Nature of variability	*No. of degrees of freedom* (n)	*Sum of squares*	*Mean square (variance)* σ^2	*Natural logarithm of mean square* $\log_e \sigma^2$
Within arrays (Component A)	36	76.39	2.12	0.7514
Between arrays (Component B)	7	151.94	21.71	3.0778
			Difference =	2.3264
			z =	1.1632

When we divide the sums of the squares by the corresponding figures defining degrees of freedom, we have comparable measures of variance. Now it appears that the variance between arrays (21.71) is distinctly greater than the variance within arrays (2.12), in disproof of the hypothe-

sis that the same forces account for the two variances. But we have a precise test to employ in determining whether these two variances are of the same degree of magnitude, within sampling limits. This is the coefficient z, which is half the difference between the natural logarithms of the two variances. In the present case, z is equal to $\frac{3.0778 - .7514}{2}$, or 1.1632.

If the forces responsible for variation within arrays were the same as those responsible for variation between arrays (that is, if our hypothesis were true), the value of z would be zero, with a sample of infinite size. The value of z we have secured is not zero. This may be proof that our hypothesis is false, or it may merely be a result of sampling fluctuations. The value of z might be zero in a given infinite population, but a random sample would be expected to yield results deviating considerably from zero. We wish now to take account of sampling fluctuations, in determining whether the result we have secured is consistent with the hypothesis that the true value of z is zero.

In determining the significance of the present results we enter Appendix Table VI with n_1 (the number of degrees of freedom corresponding to the larger variance) equal to 7 and n_2 equal to 36. Interpolating in Table VI, we find that the 1 per cent value of z corresponding to the stated values of n_1 and n_2 is .5780.[1] A value as great as this or greater

[1] It is necessary to interpolate on both scales of the z-table. First, following the procedure indicated on a preceding page, we interpolate in respect of n_2. We obtain

for $n_1 = 6$, $n_2 = 36$, the 1 per cent value of z = .6047;	1/6 = .1667
for $n_1 = 8$, $n_2 = 36$, the 1 per cent value of z = .5580;	1/8 = .1250
Δ = .0467	Δ = .0417.

We must now interpolate on the n_1 scale, since the degrees of freedom are $n_1 = 7$, $n_2 = 36$. For $1/n_1$ we have 1/7 = .1429. The difference between 1/7 and 1/8, for which we must interpolate between the given values of z, is .1429 − .1250, or .0179.

The required 1 per cent value of $z = .5580 + \left(\frac{.0179}{.0417} \times .0467\right) = .5780$.

would occur only 1 time out of 100, as a result of sampling fluctuations, if the true value of z were zero. The actual value, 1.1632, far exceeds the 1 per cent value of z. The evidence strongly indicates that z deviates from zero not because of the play of chance, but because the forces responsible for variation between arrays are of a different order from those responsible for variations within arrays. We are justified in concluding that our results are not consistent with the assumption that the true value of z is zero. The hypothesis that the forces responsible for variation between arrays are of the same character as those responsible for variation within arrays is not tenable. The results indicate the presence of a real connection between alfalfa yield and depth of irrigation water applied.

TESTING THE HYPOTHESIS OF A LINEAR RELATIONSHIP

Since it appears that there is a relationship between these two variables, it is now in order to secure an acceptable function, defining the relationship in quantitative terms. We may do this by testing, in turn, various hypotheses concerning the form of this function, until we secure one with which the observations are not inconsistent. We shall start with the hypothesis that there is a linear relationship between alfalfa yield and depth of irrigation water applied.[1]

The first step in applying the present test is to fit a straight line to the means of the eight arrays shown in Table 117. Variation among these means (component B of the total variation) reflects the presence of correlation

[1] Each hypothesis tested should be rational, acceptable on logical grounds. If we are thinking of general relationships, prevailing over the entire range of possible observation, the assumption of a straight-line relationship between alfalfa yield and amount of irrigation water applied is not tenable. For it is not to be expected that increased irrigation will increase yield without limit. In the present case we test the hypothesis of a linear relationship in order that the demonstration of procedure may be systematic and complete, although that hypothesis is not a rational one, even within the range of the present observations.

between alfalfa yield and irrigation water applied. If the correlation is perfectly linear, all these class means will fall on the straight line; all the variation between arrays will be accounted for by the hypothesis of a linear relationship. If the relationship is substantially, though not perfectly, linear, the portion of component B not accounted for by linear regression will be insignificant. If the regression is not truly linear the residue of B not accounted for (i.e., the scatter of the means of the arrays about the straight line of regression) will be too great, and some other hypothesis concerning the character of the relationship between alfalfa yield and irrigation water applied must be employed.

A straight line fitted by the method of least squares to the means of the eight arrays is shown in Fig. 82 on page 406. The equation to the line is $Y = 5.038 + .0886X$, where Y is alfalfa yield in tons per acre and X is depth of irrigation water applied, in inches. [We should note that in the fitting process the mean of each array is weighted by the number of observations in that array. This means, merely, that six points are assumed to have coördinates of 0, 3.88 (equal to those of the mean of the first array), that four points are assumed to have coördinates of 18, 6.80 (equal to those of the mean of the third array), etc.] In Table 119 on page 510 are given the values of the means of the various arrays, and the corresponding computed values, as derived from the straight line of regression.

It is clear from the graph and the table that the fit of the straight line to the means of the arrays is not perfect. The inadequacy of the fit is measured by the sum of the squared deviations of the class means from the corresponding computed values (each squared deviation being weighted by the number of observations in the given class). This sum is equal to 44.79.

This sum, to which we may refer as B_2, is one component of B, the variation between arrays. It is that portion of the variation between arrays that is not accounted for

TABLE 119

Alfalfa Yield and Depth of Irrigation Water

(Class means and values based on linear relationship $Y = 5.038 + .0886X$)

(1) *Inches of water (class)*	(2) *No. of observations*	(3) *Mean yield of class*	(4) *Estimated yield, linear relationship (tons)*	(5) *Difference between mean yield of class and estimated yield*	(6)	(7)
	f	$\bar{Y}_p$	y_c	d	d^2	fd^2
0	6	3.88	5.04	− 1.16	1.3456	8.0736
12	6	5.63	6.10	− .47	.2209	1.3254
18	4	6.80	6.63	+ .17	.0289	.1156
24	6	7.92	7.16	+ .76	.5776	3.4656
30	6	8.98	7.70	+ 1.28	1.6384	9.8304
36	6	9.27	8.23	+ 1.04	1.0816	6.4896
48	6	9.02	9.29	− .27	.0729	.4374
60	4	8.42	10.36	− 1.94	3.7636	15.0544
						44.7920

by the hypothesis of a linear relation between yield and irrigation water. The method of deriving the other component of B is shown in Table 120.

The sum 107.15, to which we may refer as B_1, is that component of the variation between arrays which is accounted for by the hypothesis of linear regression. The items in col. (3) of Table 120 differ from 7.48, the mean of all the observations, for the reason suggested by the hypothesis. They differ, on our present assumption, because with increased applications of water yield increases in a manner defined precisely by the equation $Y = 5.038 + .0886X$. The sum of these variations, 107.15, represents, on this assumption, the full effect on alfalfa yield of variations of irrigation applications.

The total of the two sums to which we have referred as B_1 and B_2 is equal to 151.94, the total variation between arrays. Working on the hypothesis that the variables

TABLE 120

Computation of Variation in Alfalfa Yield Attributable to Irrigation Differences on the Hypothesis of Linear Regression

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Inches of water	*No. of observations*	*Estimated yield, linear relationship (tons)*	*Mean yield, all observations*	*Difference between mean yield and yield estimated on linear hypothesis* $(y_e - \overline{Y})$		
	f	y_e	$\overline{Y}$	d	d^2	fd^2
0	6	5.04	7.48	− 2.44	5.9536	35.7216
12	6	6.10	7.48	− 1.38	1.9044	11.4264
18	4	6.63	7.48	− .85	.7225	2.8900
24	6	7.16	7.48	− .32	.1024	.6144
30	6	7.70	7.48	+ .22	.0484	.2904
36	6	8.23	7.48	+ .75	.5625	3.3750
48	6	9.29	7.48	+ 1.81	3.2761	19.6566
60	4	10.36	7.48	+ 2.88	8.2944	33.1776
						107.1520

with which we are dealing stand in a linear relationship, we have broken the component B of the total variation into two portions. One of these (B_1) measures the variation between arrays that is accounted for by the linear hypothesis; the other (B_2) measures the variation between arrays that is not accounted for by that hypothesis. We should expect some departure from linearity in a sample such as ours, even though it were drawn from a universe marked by a perfect linear relationship. But there are limits to the deviations that might reflect fluctuations of sampling. The question we now face is whether B_2 is small enough to be accepted as the resultant of random factors, or whether it is so large as to represent a breakdown of our hypothesis.

In our earlier discussion we noted that component A of the total variation measured the influence of a host of random forces affecting alfalfa yield, forces other than the irrigation factor. Component A, therefore, serves as an

index of the magnitude of random forces, and hence as a standard defining the probable limits of sampling fluctuations, in so far as these are present in component B. We may use component A, which relates to variation within arrays, as a yardstick in determining whether B_2 is attributable to fluctuations of sampling, or whether it is too large to be so explained.

In comparing components A and B_2 account must be taken of the number of degrees of freedom present in each. This has already been established for A. The following tabular summary of the operations just performed may help to explain the relations involved for B_2.

Nature of variability	*No. of degrees of freedom*	*Sum of squares*	*Mean square*
Between arrays, due to linear regression (Component B_1)	1	107.15	
Deviations from straight line of regression (Component B_2)	6	44.79	7.47
Total variation between arrays (Component B)	7	151.94	

The seven degrees of freedom entering into component B are divided, one to component B_1 and six to component B_2. That the points on a straight line vary from one another with one degree of freedom is clear from a consideration of a linear equation $y = a + bx$. That the values of y may differ is due to the presence of the coefficient b, which defines the slope. If b were zero, the equation would define a horizontal line, with values of y constant. It is the slope that constitutes the one degree of freedom among points defined by a linear equation. With respect to B_2, we are dealing with eight points, to which a straight line has been fitted. If there were but two points both of them would lie on the line; there would be no possibility of deviation. With three points, one degree of freedom to deviate is introduced; with eight points there are six degrees of freedom. The degrees of freedom to deviate from any

fitted curve are obviously equal to the number of points to which the curve is fitted, less the number of constants in the equation to that curve.

Dividing 44.79 by 6 we may secure, then, the value of the variance (the mean square) comparable to the variance of component A. A test of our hypothesis again reduces to a comparison of variances. This appears in Table 121.

TABLE 121

A Test of the Hypothesis of Linear Relationship

Nature of variability	*Degrees of freedom* n	*Mean square (variance)* σ^2	*Natural logarithm of mean square* $\log_e \sigma^2$
Within arrays (Component A)	36	2.12	.7514
Deviation from straight line of regression (Component B_2)	6	7.47	2.0109
		Difference =	1.2595
		z =	.6298

The variation within arrays reflects the play of random factors, independent of irrigation. The force of these factors is indicated by a variance of 2.12. If similar random factors, independent of irrigation, were responsible for the deviations of the means of the eight arrays from the straight line of regression, we should expect the variance that measures such deviations to be of the same order of magnitude. Actually it is much greater, 7.47. But we cannot say, from inspection, that the difference between the two variances is not due to fluctuations of sampling. An accurate test is needed. We may compute the coefficient z, half the difference between the natural logarithms of the two variances, and apply such a test.

From the values given we secure a value of z equal to .6298. In determining whether this value is significantly different from zero, use must be made again of Fisher's tables. For the values of n_1 and n_2 are relatively small

and unequal, and the distribution of z under these conditions would not be sufficiently close to the normal type to justify the use of its standard deviation. Entering Appendix Table VI with n_1 equal to 6, n_2 to 36, we find that the 1 per cent value of z is .6047. We take this to mean that, if the true value of z were zero, random sampling fluctuations would be expected to give a value of z as great as .6047, or greater, only one time out of 100 trials. The actual value of z in the present instance is greater than .6047. Only rarely, less frequently than one time out of 100, would chance account for a value of z as great as the one observed. We conclude, therefore, that random forces, of the type responsible for variation within arrays, are not responsible for the deviations of the means of the eight arrays from the straight line of regression. These deviations are too great to be consistent with the hypothesis that there is a linear relationship between alfalfa yield and depth of irrigation water. This equation fails to account, adequately, for the observed variation between arrays.

TESTING THE HYPOTHESIS OF A CURVILINEAR RELATIONSHIP

We may now test the hypothesis that a power curve of the second degree ($Y = a + bX + cX^2$) defines the relation between alfalfa yield and depth of irrigation water applied. The procedure is identical with that followed in the case of the straight line. By the method of least squares we determine the best values of the constants in an equation of the desired form. The curve is fitted to the means of the eight arrays, each weighted by the number of observations in that array. The derived equation is $Y = 3.539 + .2527X - .002827X^2$. The curve appears graphically in Fig. 82, and the computation of the sum of the squared deviations from it is shown in Table 122.

The inadequacy of the fit is measured this time by the figure 4.61, the sum of the squared deviations from the power curve of the second degree. This sum, to which we

TABLE 122

Alfalfa Yield and Depth of Irrigation Water

(Class means and values based on a power curve of the second degree)

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Inches of water (class)	*No. of observations*	*Mean yield of class (tons)*	*Estimated yield, from equation (tons)*	*Difference between mean yield of class and estimated yield*		
	f	$\overline{Y}_p$	y_c	$\overline{Y}_p - y_c$ d	d^2	fd^2
0	6	3.88	3.54	+ .34	.1156	.6936
12	6	5.63	6.16	− .53	.2809	1.6854
18	4	6.80	7.17	− .37	.1369	.5476
24	6	7.92	7.98	− .06	.0036	.0216
30	6	8.98	8.58	+ .40	.1600	.9600
36	6	9.27	8 97	+ .30	.0900	.5400
48	6	9.02	9.16	− .14	.0196	.1176
60	4	8.42	8.52	− .10	.0100	.0400
						4.6058

may refer as B_4, is a component of B, the variation between arrays. It is the portion that is not accounted for by the hypothesis of a curvilinear relationship, of the type assumed, between alfalfa yield and irrigation water applied. The other component of B is derived by the method indicated in Table 123 on page 516.

We may designate by B_3 the sum 147.32. This is the component of the variation between arrays that is accounted for by the hypothesis of a relationship defined by a second degree curve. The items in col. (3) of Table 123 differ from the mean of all the observations, on our present assumption, because alfalfa yield varies with increased applications of water in a manner defined by the equation

$$Y = 3.539 + .2527X - .002827X^2.$$

We have again broken B, the total variation between arrays, into two components, B_3 representing the influence

TABLE 123

Computation of Variation in Alfalfa Yield Attributable to Irrigation Differences on the Hypothesis of a Non-Linear Regression

(1) *Inches of water*	(2) *No. of observations* f	(3) *Estimated yield, equation of second degree* y_c	(4) *Mean yield, all observations* $\overline{Y}$	(5) $y_c - \overline{Y}$ d	(6) d^2	(7) fd^2
0	6	3.54	7.48	− 3.94	15.5236	93.1416
12	6	6.16	7.48	− 1.32	1.7424	10.4544
18	4	7.17	7.48	− .31	.0961	.3844
24	6	7.98	7.48	+ .50	.2500	1.5000
30	6	8.58	7.48	+ 1.10	1.2100	7.2600
36	6	8.97	7.48	+ 1.49	2.2201	13.3206
48	6	9.16	7.48	+ 1.68	2.8224	16.9344
60	4	8.52	7.48	+ 1.04	1.0816	4.3264
						147.3218

of the irrigation factor, working in accordance with a definite law, and B_4 representing random factors, or random factors combined with the irrigation factor. (The irrigation factor enters into B_4 to the extent that the hypothesis in question fails to take account of the true relation between alfalfa yield and depth of water applied.) This is, of course, a different division of B from that resulting from the application of a linear hypothesis. The present division may be set down in summary.

Nature of variability	*No. of degrees of freedom*	*Sum of squares*	*Mean square*
Between arrays, due to regression of second degree (Component B_3)	2	147.32	
Deviations from second degree curve of regression (Component B_4)	5	4.61	.92
Total variation between arrays (Component B)	7	151.93	

The seven degrees of freedom entering into component B are now divided, five to component B_4 and two to compo-

nent B_3. The reasons for this allocation of the degrees of freedom are similar to those presented in discussing the linear hypothesis. As regards component B_3, the item now of chief concern to us, it is clear that when a curve defined by an equation with three constants is fitted to eight points there are five degrees of freedom to deviate from that curve.

Dividing 4.61 by 5 we secure .92, the value of the variance comparable to the variance of component A. For again we must use a criterion based on A, in determining the limits within which variation due to random factors, independent of irrigation, may play. We come again to a comparison of variances.

TABLE 124

A Test of the Hypothesis of Curvilinear Relationship

Nature of variability	*Degrees of freedom* n	*Mean square (variance)* σ^2	*Natural logarithm of mean square* $\log_e \sigma^2$
Within arrays (Component A)	36	2.12	.7514
Deviation from second degree curve of regression (Component B_4)	5	.92	− .0834
		Difference =	− .8348
		z =	− .4174

In this case the degree of deviation from the curve of regression defined by the power curve of the second degree is actually less than the deviation within arrays, which serves as our yardstick. The value of z is therefore negative, equal to − .4174. This measure may be tested for significance by the methods previously discussed. The z-table is entered with $n_1 = 36$ (the number corresponding to the larger of the two variances), $n_2 = 5$. Interpolating in the table for these values we obtain 1.1158 as the 1 per cent value of z. The present value is distinctly less than this. The difference between the two measures of variance is

not significant. The departures from the curve of regression may be attributed to "chance," that is, to random factors independent of the irrigation factor.

In following this general procedure it is necessary to test different hypotheses (i.e., different functions) only until the difference between the variance defined by component A and the variance defining departures from the curve of regression be small enough to be attributed to the play of chance. Thus, if a P of .05 constitutes our standard, the difference between the two variances given in the preceding table, as measured by z, might be positive and as great as .4536, without leading to rejection of the hypothesis being tested. It could be as great as .6370 if our standard of significance were a P of .01.[1] A rather exceptionally close fit by the second degree curve we have employed gives us the negative value of z we have actually obtained.

We have arrived, then, at an hypothesis concerning the relation between alfalfa yield and depth of irrigation water applied, with which observed facts are not inconsistent. Our observations, be it noted, do not establish the truth of this hypothesis. Other hypotheses might be equally tenable, and perhaps even more closely in accord with the facts.[2]

[1] These figures are derived from Tables VI and VII by the process of interpolation described above, with $n_1 = 5$ and $n_2 = 36$. (n_1 is taken as equal to 5, of course, only when B_4 is greater than A; n_1 is always taken to represent the number of degrees of freedom corresponding to the larger of the two variances being compared.) This method of interpolation is applicable over the range of the z-table, except for the corner relating to values of n_1 in excess of 24 and values of n_2 in excess of 30. For dealing with cases in this region, R. A. Fisher gives the following formulas for approximating the desired quantities:

$$5 \text{ per cent value of } z = \frac{1.6449}{\sqrt{h-1}} - .7843\left(\frac{1}{n_1} - \frac{1}{n_2}\right)$$

$$1 \text{ per cent value of } z = \frac{2.3263}{\sqrt{h-1}} - 1.235\left(\frac{1}{n_1} - \frac{1}{n_2}\right)$$

In these formulas h is the harmonic mean of n_1 and n_2. That is

$$\frac{1}{h} = \frac{\frac{1}{n_1} + \frac{1}{n_2}}{2}.$$

[2] We could, of course, fit a curve of still higher degree, the equation to which

All that we can say is that the observed facts do not disprove the hypothesis. If the hypothesis is tenable on rational grounds, we have reached a conclusion upon which we may rest, for the time.

SUMMARY: VARIANCE ANALYSIS IN THE MEASURE OF RELATIONSHIP

The procedure employed in the last example may be summarized and certain measurements presented which show the relation of this procedure to methods discussed in preceding chapters. The quantitative results are presented in Table 125.

TABLE 125

Component Elements of the Variability of Alfalfa Yield, and Various Measures of Correlation

Total variability of observations relating to alfalfa yield, and components of this total		*Test of significance*	*Measure of correlation*
Total variability (sum of squared deviations from Mean)	228.33		
1. Division of total variability into:			
A. Variation unrelated to irrigation factor (i.e., variation within arrays)	76.39		
B. Variation attributable to irrigation factor, and to other causes, in indeterminate proportions (i.e. variation between arrays)	151.94	$z = 1.1632$ 1 per cent value of z $= .5780$	Correlation ratio $\eta^2 = \frac{151.94}{228.33}$ $= .6654$ $\eta = .82$

(*Footnote 2 continued from page 518.*)
contained four constants, or more, instead of the three constants in the equation actually employed. The deviations from this curve of higher degree would be smaller than from the curve of second degree, and z would be correspondingly smaller. It is a principle of scientific procedure, however, to employ the simplest acceptable function. Needless complexities, whether in the form of unnecessary assumptions or of unnecessary constants in an equation of relationship, are rigorously avoided.

TABLE 125—*Continued*

Component Elements of the Variability of Alfalfa Yield, and Various Measures of Correlation

Total variability of observations relating to alfalfa yield, and components of this total		*Test of significance*	*Measure of correlation*
2. Division of component B of (1) above into:			
B_1. Variation attributable to irrigation factor on the assumption of a linear relationship	107.15		
B_2. Variation attributable to irrigation factor, but not explainable in terms of a linear relationship, and to other causes, in indeterminate proportions	44.79	$z = .6298$ 1 per cent value of z $= .6047$	Coefficient of correlation $r^2 = \frac{107.15}{228.33}$ $= .4693$ $r = .69$
3. Division of component B of (1) above into:			
B_3. Variation attributable to irrigation factor on the assumption of a relationship defined by power curve of second degree	147.32		
B_4. Variation attributable to irrigation factor, but not explainable in terms of power curve of second degree, and to other causes, in indeterminate proportions	4.61	$z = -.4174$ 1 per cent value of z $= 1.1158$	Index of correlation $\rho^2 = \frac{147.32}{228.33}$ $= .6452$ $\rho = .80$

The meaning of this summary should be clear, with reference to the preceding demonstration. Component A of the total variability, being independent of the influence of the irrigation factor, is the yardstick, or standard of reference, which is used in all the tests of significance noted in the second column. Component B, in the first test,

is shown to be clearly greater than A, when account is taken of the number of degrees of freedom present in the two quantities. Thereafter, component B is broken into sub-components, first on the hypothesis that alfalfa yield and irrigation are related by a linear function, next on the hypothesis that the relationship is defined by a power curve of the second degree. The evidence is not consistent with the first of these hypotheses, and it is rejected. (The hypothesis would be rejected on rational grounds, as well as on the basis of empirical evidence.) The results are not inconsistent with the second hypothesis, and we may accept it, subject to the possibility of modification on the basis of later experience.

Three abstract measures of degree of correlation between alfalfa yield and applications of irrigation water are given in the right-hand column. All of these may be derived directly from the quantities employed in the variance analysis. Study of the elements of these correlation measures, and of the relation of the several measures to the corresponding hypothesis, will provide a suggestive review of the general problem of correlation.

We should note here that an assumption of normality is implied in the comparison of standard deviations, or of variances, in this type of analysis. Minor departures from normality do not materially affect the procedure, but substantial departures do so. The conversion to other forms (such as logarithms or reciprocals) of observations not normally distributed in natural terms will sometimes yield normal distributions. Where this is possible, the precision of the method of variance analysis is increased by such conversion. Limitations arising out of material departures from normality may be avoided, also, by the use of ranks, as is done in the computation of the coefficient of rank correlation. Appropriate procedures have been developed by Milton Friedman.[1]

[1] "The Use of Ranks to Avoid the Assumption of Normality Implicit in the

Variance Analysis in Testing the Significance of Seasonal Fluctuations

The methods outlined in this chapter are applicable to certain of the problems encountered in the analysis of time series. They are peculiarly appropriate in determining whether the seasonal fluctuations observed in a given series represent a true seasonal pattern. Apparent seasonal movements would be present in any series of observations covering a period of years, by months. Chance factors would create some differences between averages of all the Januaries, all the Februaries, etc., even though no true seasonal movement existed. We require an objective test, to be used in determining whether the differences among such monthly averages are significant or not.

The entries in the body of Table 126 are the figures obtained when freight car loadings by months, for the period 1918–1927, are expressed as percentages of linear trend values. (The original data are given in Chapter VIII.) The arithmetic mean of the ten items for January appears in the bottom row, with similar means for the other months.[1] The test for seasonality involves answering the question: Do these means differ significantly from 99.9867, the average value, of the 120 items in the table? In seeking to answer this question we must break the total variance of the freight car loadings data into its elements. We wish to define that portion of the total variance apparently due to seasonal movements. This may then be appraised with reference to a yardstick representing what we may call the residual variability of the series.

In computing the total variance we may make use of the

Analysis of Variance," *Journal of the American Statistical Association*, Vol. 32, Dec. 1937, 675–701.

[1] These means, it may be noted, are not precisely the same as the seasonal indices given in Chapter VIII. In seeking to improve the representativeness of the monthly indices, only the four central items for each month were used in the averaging process employed in that chapter. Here it is necessary to employ the arithmetic mean of all the items for each month.

TABLE 126

Freight Car Loadings in the United States, 1918–1927

Monthly Values as Percentages of Trend, with Computations Required in the Analysis of Variance

1 Year	2 Jan.	3 Feb.	4 Mar.	5 Apr.	6 May	7 June	8 July	9 Aug.	10 Sept.	11 Oct.	12 Nov.	13 Dec.	14 Sum.	15 Mean	16 Sum of squares
1918	83.8	96.0	107.1	110.8	113.5	115.9	122.2	120.9	119.8	115.6	102.1	89.4	1,297.1	108.092	141,982.37
1919	90.3	85.0	85.9	88.8	93.4	97.7	108.6	108.9	116.9	117.5	97.8	91.7	1,182.5	98.542	118,010.95
1920	98.9	93.4	101.8	87.4	103.0	106.8	107.1	114.9	114.4	118.7	104.1	88.7	1,239.2	103.267	129,075.98
1921	82.8	80.1	80.6	82.2	88 3	88.3	86.7	93.4	96.9	107.0	87.5	77.8	1,051.6	87.633	92,906.98
1922	79.4	86.1	92.8	81.1	87.7	93.7	91.5	95.7	103.6	109.3	106.7	92.5	1,120.1	93.342	105,528.93
1923	94.2	94.7	101.3	103.9	107.4	111.2	108.0	114.0	113.2	116.6	104.8	89.7	1,259.0	104.917	132,897.76
1924	93.0	98.1	98.7	94.0	96.0	97.0	94.1	103.3	110.3	115.8	103.4	91 9	1,195.6	99.633	119,713.90
1925	94.0	95.4	97.3	97.7	101.6	103.5	101.6	111.5	111.4	114.6	105.8	95.4	1,229.8	102.483	126,577.28
1926	94.7	95.7	98.4	98.8	103.9	107.1	105.4	112.2	115.5	119.7	105.1	86.4	1,242.9	103.575	129,713.51
1927	94.9	95.9	99.9	96.8	100.1	101.6	97.1	106.3	108.1	108.7	91.2	80.0	1,180.6	98.383	116,842.48
Sum.	906.0	920.4	963.8	941.5	994.9	1,022.8	1,022.3	1,081.1	1,110.1	1,143.5	1,008.5	883.5	1,199.8400		1,213,250.14
Mean	90.60	92.04	96.38	94.15	99.49	102.28	102.23	108.11	111.01	114.35	100.85	88.35		99.9867	

familiar relation $\frac{\Sigma d^2}{N} = \frac{\Sigma(d')^2}{N} - c^2$ where d is the deviation of an observation from the true mean, d' is the deviation from an assumed mean, and c is the difference between true and assumed means. In this case we take the assumed mean at 0 on the original scale, and c is thus equal to the mean. Since we wish to work with sums of squared values, we use the relationship

$$\Sigma d^2 = \Sigma(d')^2 - Nc^2.$$

(The mean should be computed to more places than are to be finally retained, since the process of squaring and multiplying by N greatly magnifies even slight errors.)

The entries in col. (16) of Table 126 are the sums of the squares of the items in the body of the table. Inserting the proper values in the above formula, we have

$$\begin{aligned}\Sigma d^2 &= 1{,}213{,}250.14 - 120 \times 99.9867^2 \\ &= 1{,}213{,}250.14 - 1{,}199{,}680.82 \\ &= 13{,}569.32.\end{aligned}$$

As in the alfalfa problem discussed above, this total may be broken into an element representing variance between the monthly means and variance within the several months. (Reference here is to the *columns* of Table 126.) The variance between months may be computed directly from the monthly means.

Thus:

Sum of squares of deviations of monthly means from grand mean

$$\begin{aligned}&= 10 \times (99.9867 - 90.60)^2 + 10 \times (99.9867 - 92.04)^2 \\ &\quad + 10 \times (99.9867 - 96.38)^2 + \ldots \\ &\quad + 10 \times (99.9867 - 88.35)^2.\end{aligned}$$

That is, the deviation of each monthly mean from the grand mean is squared and weighted by the number of items represented by that mean; the sum of the twelve figures thus obtained is the required measure of variability between months.

An alternative shorter method may be employed in determining the variance between months, utilizing the relationship

$$\Sigma d^2 = \Sigma(d')^2 - Nc^2.$$

Here each d' is the mean value for a given month. Each squared value must be weighted by the number of items represented by the mean. Thus

$$\Sigma(d')^2 = 10(90.60)^2 + 10(92.04)^2 + 10(96.38)^2 + \ldots + 10(88.35)^2$$
$$= 1{,}207{,}068.40.$$

The correction factor, Nc^2, is the same as in the first operation. We have, then,

Sum of squares of deviations of monthly means from grand mean $= 1{,}207{,}068.40 - 1{,}199{,}680.82$
$= 7{,}387.58.$

This sum measures that portion of the total variability that may be attributed to seasonal fluctuations. Is it significant or does it merely reflect the play of the mass of undifferentiated factors we call chance?

In answering a similar question concerning the alfalfa problem we used as a yardstick the variability independent of the one factor the effects of which were being studied — namely, irrigation differences. In the present case we could obtain a measure that is independent of seasonality by computing the variability *within* the several columns of Table 126. That is, each item in col. (2) could be deducted from the January mean, 90.60, and the sum of the squared deviations in this column obtained; a similar sum could be obtained for each of the other columns numbered from (3) to (13). The grand sum of these figures would be the variability within arrays — variability clearly independent of seasonal forces since only differences among items for the same month enter into it. This sum, measuring variability within columns, has a value of 6,181.74. The variability between columns plus the variability within

columns is, of course, equal to the total. That is, 7,387.58 + 6,181.74 = 13,569.32.

The measure of variability within columns will not serve in the present instance, however. The yardstick should be a measure of the variability due to "chance"—to the play of a mass of random factors which may not be observed and measured individually. Effects that can be clearly attributed to specific causal forces should not be included in the yardstick. But some of the variability within months may be clearly assigned to changes associated with the classification by years. The average of the 12 monthly items for 1918 is 108.09; that for the 12 months in 1921 is 87.63. The former was a year of prosperity, the latter one of depression. Clearly, some of the differences among the items in the January column, or in the May column, are definitely attributable to cyclical forces that raise all the monthly figures for one year and depress all the monthly figures for another year. (The influence of trend is not present, since the items in the body of the table are actual values expressed as percentages of trend.) The variability within months should be corrected by the subtraction of that portion of it that may be attributed to factors affecting yearly conditions as a whole.

The influence of cyclical and other forces affecting whole years is measured by differences between the averages for 1918, 1919, 1920, and the other years covered. These averages are given in col. (15) of Table 126. The desired quantity may be obtained by the precise methods used in measuring the variability between months. We have

$$\Sigma d^2 = \Sigma(d')^2 - Nc^2$$

or,

Sum of squares of deviations of yearly averages from grand mean

$$= (12 \times 108.092^2 + 12 \times 98.542^2 + 12 \times 103.267^2 + \ldots + 12 \times 98.383^2) - 120 \times 99.9867^2$$

$$= 1{,}203{,}537.88 - 1{,}199{,}680.82$$

$$= 3{,}857.06.$$

(There will, of course, be ten squared items within parentheses, one for each of the ten years covered by the data.) Subtracting 3,857.06 from 6,181.74, the measure of total variability within the columns, we have 2,324.68 as the balance. This is the desired yardstick. It measures that portion of the variability among the original items which is clearly independent of the seasonal influence. Secondly, it has been corrected by the subtraction of that portion of the variability within months which is attributable to cyclical and other factors responsible for broad changes from year to year. The final balance represents the play of forces independent both of seasonal movements and of broad swings affecting each yearly value as a whole. This *residual variability*, measured by the figure 2,324.68, reflects the play of all those random, undifferentiated factors we lump together as chance.[1]

This residual variability may be most readily computed by subtracting from the total variability the two figures measuring, respectively, variability between the means of the months and variability between the means of the years. At this stage of the computation these figures will be in the form of sums of squared deviations. The form of organization employed in Table 126 on page 528 is convenient for these calculations.

In the application of the test of significance, account must be taken of the number of degrees of freedom entering into each of these measures of variability. Table 127 indicates a suitable procedure.

[1] When, as in the present example, the influences of the two variables, or principles of classification, are independent, it is valid to use the residual variability thus computed as a measure of the strength of random factors. If these influences are not independent (if, in terms of the above example, seasonal movements affecting the monthly averages and cyclical movements affecting the annual averages should be correlated), the residual quantity will not be an accurate measure of truly random factors. When the residual quantity which is used as the yardstick in variance analysis is derived from observations that are alike in respect of *both* principles of classification (i.e., when the quantity measures variance within cells obtained by the application of a two-fold principle of classification) this difficulty does not arise. An example of this type is given in Appendix E.

TABLE 127

Analysis of Variance of Freight Car Loadings and Test of Seasonality

(1) *Nature of variability*	(2) *No. of degrees of freedom* (n)	(3) *Sum of squares*	(4) *Mean square (variance)* σ^2 (3) ÷ (2)	(5) *Natural logarithm of mean square* $\log_e \sigma^2$
Between means of years	9	3,857.06		
Between means of months	11	7,387.58	671.598	6.50970
Residual variability	99	2,324.68	23.482	3.15627
Total	119	13,569.32	Difference =	3.35343
			$z = \frac{3.35343}{2}$	
			=	1.67671

The item 3,857.06 measures the degree of difference between 10 yearly averages. Nine degrees of freedom are represented in this figure. (The use of weights in computing the sum of the squares does not affect the number of degrees of freedom.) Similarly, 11 degrees of freedom are represented in the measure of variability between the 12 monthly means. The total variability is computed from 120 items, so there are 119 degrees of freedom in all. The number of degrees of freedom in the residual variability is, therefore, 119 − (11 + 9), or 99.

The variance between the means of months (i.e., the mean square) is 671.598. The residual variance is 23.482. The test of seasonality reduces to the question: May the variance between the means of months be attributed to the random forces responsible for the residual variance? Unless the variance between the monthly means is significantly greater than the residual variance, no significance may be attached to the observed differences between the averages for January, February, March, and the other months.

The test is applied with reference to the measure z, which is equal to half the difference between the natural logarithms of the two variances being compared. From the entries in Table 127 we compute z as equal to 1.67671. Referring to Appendix Table VI we find that for $n_1 = 11$ and $n_2 = 99$, the 1 per cent value of z is approximately .44. The present value is distinctly greater than this. The results are not consistent with the hypothesis that the true value of z is zero. There is clear evidence of the existence of a definite seasonal pattern in freight car loadings.[1]

The same yardstick may be applied in testing whether the differences between the yearly averages are significant. The rather wide variations from year to year in the average values of the items in the body of Table 126 represent, presumably, the play of cyclical forces plus major "accidental fluctuations" affecting yearly totals. (The trend factor, had it not been eliminated, would have combined with these other two to create differences among the yearly totals.) But are these year-to-year differences great enough to be attributed to definite forces other than the chance factors represented in the residual variance we are using as yardstick?

The variance between means of years is equal to 3,857.06 ÷ 9, or 428.562. Is this significantly greater than 23.482, the residual variance? Following the procedure illustrated in Table 127 we obtain 1.35352 as the value of z. The 1 per cent value of z, for $n_1 = 9$, $n_2 = 99$, is approximately .47. The test indicates that the differences between the annual averages are due to definite forces other than the random factors represented in the residual variance.

[1] In the test here applied we are proceeding on the assumption that the seasonal pattern is constant from year to year. If it is not constant, the accuracy of the residual variability, as a measure of "chance" factors, and of the measure of variability between months will be affected, and the significance of the results will be lessened. If there is reason to believe that seasonal movements have changed over the period covered, tests of the kind suggested in Chapter VIII should precede the tests here discussed.

REFERENCES

Fisher, R. A., *Statistical Methods for Research Workers.* Chap. 8.

Fisher, R. A., *The Design of Experiments.* Chap. 4.

Friedman, Milton, "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *Journal of the American Statistical Association,* Dec. 1937.

Schultz, T. W. and Snedecor, G. W., "Analysis of Variance as an Effective Method of Handling the Time Element in Certain Economic Statistics," *Journal of the American Statistical Association,* March 1933.

Snedecor, G. W., *Calculation and Interpretation of Analysis of Variance and Covariance.*

Snedecor, G. W., *Statistical Methods.* Chaps. 10–12, 15.

Tippett, L. H. C., *The Methods of Statistics.* Chap. 6.

Yule, G. U. and Kendall, M. G., *An Introduction to the Theory of Statistics.* Chap. 23.

CHAPTER XVI

THE MEASUREMENT OF RELATIONSHIP: MULTIPLE AND PARTIAL CORRELATION

In dealing with methods of defining correlation in the preceding chapters we have been concerned with problems involving only two variables, a dependent variable and a single independent variable. We have found, in certain cases, a fairly high degree of correlation between the two variables studied. But it is obvious that, in general, economic phenomena are affected by more than one factor, that the fluctuations in a single variable may be due to the interaction of many forces. In dealing with just two variables all other factors are ignored, on the assumption, usually, that in the single independent variable are found the most important causes[1] of fluctuations in the dependent variable. Thus, in the alfalfa example given, the effect upon yield of but a single factor, irrigation, was studied. Yet variations in rainfall and temperature must have affected the yield in the different years studied. Similarly, variations in practically every factor dealt with in economic analysis are traceable to more than one cause. If our analysis is to be complete we must employ methods which will enable more than two variables to be handled at a time. We need instruments that will assist us in measuring the combined effect upon a single variable of a number of factors. Such instruments may be secured by a simple extension of methods already familiar.

In Table 128 are presented figures showing the yield of corn, per acre, in Kansas from 1890 to 1933, together

[1] This should not be taken to mean that the coefficient of correlation measures or establishes causal relationships.

TABLE 128

Corn Yield and Temperature in Kansas, 1890–1933 [1]

(1) Year	(2) Average yield per acre, in bushels X_1	(3) Average June temperature X_2	(4) Average July temperature X_3	(5) Average August temperature X_4
1890	15.6	77.6	83.1	76.1
1891	26.7	70.7	74.0	75.1
1892	24.5	73.4	77.5	76.5
1893	21.3	74.7	79.5	73.8
1894	11.2	74.2	77.8	78.0
1895	24.3	71.7	74.9	76.0
1896	28.0	74.1	78.1	78.7
1897	18.0	76.6	80.2	76.0
1898	16.0	75.0	77.7	78.2
1899	27.0	73.9	76.2	80.6
1900	19.0	74.9	77.9	81.0
1901	7.8	77.3	85.0	79.1
1902	29.9	70.9	76.8	78.2
1903	25.6	67.2	78.3	75.3
1904	20.9	70.4	75.6	74.6
1905	27.7	75.5	74.5	78.7
1906	28.9	71.8	73.8	76.3
1907	22.1	72.0	78.4	78.1
1908	22.0	72.1	75.8	76.2
1909	19.9	73.1	78.1	80.1
1910	19.0	72.2	79.5	75.7
1911	14.5	80.5	78.6	76.4
1912	23.0	69.3	79.9	77.4
1913	3.2	74.2	82.1	84.2
1914	18 5	78.2	79.9	78.2
1915	31.0	69.2	74.0	70.1
1916	10.0	70.3	81.2	79.6
1917	13.0	72.8	80.8	73.4
1918	7.1	78.4	78.3	82.3
1919	15.2	72.3	80.2	78.3
1920	26 5	72 8	77.6	72.9
1921	22.2	74.4	79.2	78.6
1922	19.3	75.2	77.0	80.1
1923	21.7	73.3	79.4	78.3
1924	21.7	74.3	75.1	79.0
1925	16.6	77.7	79.7	77.4
1926	11.0	72.5	78.4	79.1
1927	30.0	70.9	76.9	73.1
1928	27.0	67.7	78.1	77.1
1929	17.5	72.2	78.8	78.9
1930	12.0	73.1	81.7	80.3
1931	17.5	78.1	80.6	76.1
1932	18.5	74.3	81.8	79.2
1933	11.5	80.5	81.4	76.8

[1] The data of corn yield are from *Bulletin* 515, U. S. D. A., and from the *Yearbooks* of the U. S. D. A. Temperature data are from reports of the U. S. Weather Bureau for Dodge City, Concordia and Iola.

with the average June, July, and August temperatures for each of these years.

The Relation between Corn Yield and Temperature: Preliminary Analysis

It is known that corn yield is affected by the temperature during the growing season. The object of the present study is the determination of the precise relation between yield and temperature during each of the three months given, in order to secure a basis for estimating the yield from a knowledge of the temperature. As certain growing months are more important than others, the relation between temperature and yield may be determined, first, for each of the three months separately.

The equation which describes the relationship between yield per acre and June temperature will be of the type

$$X_1 = a + b_{12}X_2.$$

The equation describing the relationship between yield per acre and July temperature will be of the type

$$X_1 = a + b_{13}X_3.$$

(In each case X_1 represents average corn yield per acre, for the State, while X_2, X_3, etc., represent the absolute temperature, in degrees Fahrenheit.) Instead of using to represent the variables the symbols Y and X, as in the preceding examples, X_1, X_2, X_3, etc., are employed, X_1 representing in this case the dependent variable. The symbol for the constant representing the slope (the coefficient of regression) is, in the first instance above, b_{12}. The subscripts 1 and 2 indicate the variables to which this constant refers, the first subscript always representing the dependent variable (X_1 in the example cited), the second the independent variable (X_2 in the illustration above). These subscripts are necessary to distinguish the different constants when several variables enter into the problem.

The meaning is precisely the same as in the former examples when no subscripts were needed because only two variables were dealt with.

Solving the proper normal equations for the constants in the equation which describes the average relationship between yield per acre and June temperature, we have

$$X_1 = 100.35 - 1.096X_2.$$

The value of S_{12} may be determined from the formula

$$S^2_{12} = \frac{\Sigma(X_1^2) - a\Sigma(X_1) - b_{12}\Sigma(X_1X_2)}{N}.$$

(The subscripts to S, and those to r which appear below, have the same meaning as those employed in the preceding paragraph.) Substituting the given values, and solving, we have

$$S^2_{12} = 33.593$$

and

$$S_{12} = 5.80.$$

The significance of the standard error, S, as a measure of the reliability of estimates based upon the equation of relationship, has been fully explained. In judging of the usefulness of the equation, S_{12} should be compared with σ_1 (the standard deviation of X_1) which may be looked upon as a measure of the reliability of estimates based upon the arithmetic mean of the variable X_1. For this we have

$$\sigma_1 = 6.68.$$

Clearly, the estimates from the equation are more reliable than those based upon the mean. The coefficient of correlation, r, expresses this relationship in abstract terms. We may get this value from the equation

$$r^2_{12} = \frac{a\Sigma(X_1) + b_{12}\Sigma(X_1X_2) - Nc_1^2}{\Sigma(X_1^2) - Nc_1^2}.$$

Solving for r, and giving it the sign of b_{12}, we have

$$r_{12} = -.4984.$$

These values indicate a negative correlation, though not a high one, between yield per acre of corn and June temperature in Kansas. Let us see if the estimates could be improved if based upon the temperature in July instead of in June.

The values needed in this study may be computed from Table 128. Solving for the constants in the equation of regression, we secure the equation

$$X_1 = 166.07 - 1.866X_3.$$

For the standard error, we have

$$S_{13} = 4.81$$

and for the coefficient of correlation

$$r_{13} = -.6948.$$

We have here a closer relation and a better basis for estimates than in the case when June temperature was considered.

Repeating the process for yield per acre and August temperature, we have

$$X_1 = 119.45 - 1.288X_4$$
$$S_{14} = 5.78$$
$$r_{14} = -.5013.$$

August temperature, it is evident, also affects the corn yield in Kansas, a low temperature conducing to yield above normal. The relationship is not so close as in the case of July temperature, but it is still significant. What is needed now is some method of combining these three factors, in order that an estimate may be based upon a knowledge of their influence, in combination, upon the yield of corn. The addition or averaging of the temperatures in the three months will not do, for July is obviously more important than either of the other months. The principle of the method by which this may be accomplished is simple.

The Estimation of Corn Yield from Three Independent Variables

The estimating or regression equation in the present case will be one in which there is a single dependent variable (corn yield) and three independent variables. It will be of the form

$$X_1 = a + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4.$$

If we can determine the values of the four constants, we may substitute given values of X_2, X_3, and X_4 in the equation and thus get an estimate for X_1 in precisely the same way as when two variables are dealt with. The method of least squares affords the means of solving for the required constants.

The symbols require a word of explanation, as a perfectly simple equation is given a rather ponderous appearance by all the subscripts employed. The symbol b_{12}, it has been explained, represents the coefficient of regression of X_1 on X_2 (i.e., the slope of the line describing their relationship, X_1 being dependent) when these two variables alone are included in the study. The symbol $b_{12.34}$ represents the *coefficient of net regression* of X_1 on X_2. The addition of the subscripts 3 and 4 to the right of the period means, simply, that the variables X_3 and X_4 have been included in the study and the effects of their variations eliminated, in so far as this one constant ($b_{12.34}$) is concerned. This constant measures the weight which must be given to the variable X_2 in an estimate of X_1 based upon the three independent variables, X_2, X_3, and X_4. It will not, of course, be the same as b_{12}, which indicates the weight given to X_2 when an estimate of X_1 is based upon X_2 alone. Similarly the constant $b_{13.24}$, the coefficient of net regression of X_1 on X_3, measures the weight given to X_3 when X_2 and X_4 are also included. Each coefficient represents a single, simple constant, but the subscripts are necessary in order that the

precise meaning of this constant may be clear. The subscripts to the left of the period are termed *primary subscripts*, those to the right *secondary subscripts*.

FORMATION AND SOLUTION OF THE NORMAL EQUATIONS

The first task[1] is the securing of the normal equations required in solving for the constants in the estimating equation given above. Following the usual procedure[2] we have:

$$\begin{aligned}
\text{I}\quad & \Sigma(X_1) = Na + b_{12.34}\Sigma(X_2) + b_{13.24}\Sigma(X_3) + b_{14.23}\Sigma(X_4)\\
\text{II}\quad & \Sigma(X_1X_2) = a\Sigma(X_2) + b_{12.34}\Sigma(X_2^2) + b_{13.24}\Sigma(X_2X_3)\\
& \qquad + b_{14.23}\Sigma(X_2X_4)\\
\text{III}\quad & \Sigma(X_1X_3) = a\Sigma(X_3) + b_{12.34}\Sigma(X_2X_3) + b_{13.24}\Sigma(X_3^2)\\
& \qquad + b_{14.23}\Sigma(X_3X_4)\\
\text{IV}\quad & \Sigma(X_1X_4) = a\Sigma(X_4) + b_{12.34}\Sigma(X_2X_4) + b_{13.24}\Sigma(X_3X_4)\\
& \qquad + b_{14.23}\Sigma(X_4^2).
\end{aligned}$$

The given values might be substituted in these simultaneous equations and solutions secured directly for the four constants. It is possible to reduce the number of normal equations by one, however, and thus lessen materially the labor of computation. This is done by using deviations from the arithmetic mean for each variable instead of absolute values, getting rid in this way of the constant term a in the original equation.

If we let A_1, A_2, A_3, etc., represent the arithmetic means of the different variables while x_1, x_2, x_3, etc., represent deviations from the means, we may replace the absolute numbers X_1, X_2, X_3, etc., by their equivalents, $x_1 + A_1$, $x_2 + A_2$, $x_3 + A_3$, etc. Making these substitutions in the normal equations, certain algebraic simplifications are pos-

[1] The approach to the problem of multiple correlation which is here taken follows that of H. R. Tolley and M. J. B. Ezekiel "A Method of Handling Multiple Correlation Problems," *Journal of the American Statistical Association*, December, 1923, 993–1003.

[2] Cf. Appendix A for a discussion of this procedure and of the methods employed in simplifying the normal equations.

sible which eliminate the first of the normal equations, and reduce the others to the following form:

$$\frac{\Sigma(x_1x_2)}{N} = \frac{\Sigma(x_2^2)}{N} b_{12.34} + \frac{\Sigma(x_2x_3)}{N} b_{13.24} + \frac{\Sigma(x_2x_4)}{N} b_{14.23}$$

$$\frac{\Sigma(x_1x_3)}{N} = \frac{\Sigma(x_2x_3)}{N} b_{12.34} + \frac{\Sigma(x_3^2)}{N} b_{13.24} + \frac{\Sigma(x_3x_4)}{N} b_{14.23}$$

$$\frac{\Sigma(x_1x_4)}{N} = \frac{\Sigma(x_2x_4)}{N} b_{12.34} + \frac{\Sigma(x_3x_4)}{N} b_{13.24} + \frac{\Sigma(x_4^2)}{N} b_{14.23}.$$

All the variables in the above equations refer to deviations from the respective arithmetic means. Therefore $\frac{\Sigma(x_1x_2)}{N}$ is simply the mean product of the variables x_1 and x_2, $\frac{\Sigma(x_2^2)}{N}$ is σ_2^2, etc. Representing the various mean products by the symbols p_{12}, p_{13}, etc., and inserting the symbols for the squares of the standard deviations, we secure, for the normal equations:

$$p_{12} = \sigma_2^2 b_{12.34} + p_{23} b_{13.24} + p_{24} b_{14.23}$$
$$p_{13} = p_{23} b_{12.34} + \sigma_3^2 b_{13.24} + p_{34} b_{14.23}$$
$$p_{14} = p_{24} b_{12.34} + p_{34} b_{13.24} + \sigma_4^2 b_{14.23}.$$

This is the most convenient form for the solution of the normal equations.

From the data, as arranged in Table 128, the following values are derived:

$$\Sigma(X_1) = 863.9 \qquad \Sigma(X_1^2) = 18{,}928.17$$
$$\Sigma(X_2) = 3{,}241.5 \qquad \Sigma(X_2^2) = 239{,}209.57$$
$$\Sigma(X_3) = 3{,}453.4 \qquad \Sigma(X_3^2) = 271{,}317.92$$
$$\Sigma(X_4) = 3{,}409.1 \qquad \Sigma(X_4^2) = 264{,}433.19$$

$$\Sigma(X_1X_2) = 63{,}198.42$$
$$\Sigma(X_1X_3) = 67{,}295.48$$
$$\Sigma(X_1X_4) = 66{,}550.84$$
$$\Sigma(X_2X_3) = 254{,}544.98$$
$$\Sigma(X_2X_4) = 251{,}246.54$$
$$\Sigma(X_3X_4) = 267{,}649.61$$

$$c_1 = \frac{\Sigma(X_1)}{N}$$

$$= 19.6341 \qquad c_1^2 = 385.4979$$

$$c_2 = 73.6705 \qquad c_2^2 = 5{,}427.3426$$

$$c_3 = 78.4864 \qquad c_3^2 = 6{,}160.1150$$

$$c_4 = 77.4795 \qquad c_4^2 = 6{,}003.0729$$

From these values, the quantities necessary for the solution of the normal equations may be readily determined. These quantities are brought together below:

$$\sigma_1^2 = \frac{\Sigma(X_1^2)}{N} - c_1^2$$

$$= \frac{18{,}928.17}{44} - 385.4979 = 44.6878$$

$$\sigma_2^2 = \frac{239{,}209.57}{44} - 5{,}427.3426 = 9.2385$$

$$\sigma_3^2 = \frac{271{,}317.92}{44} - 6{,}160.1150 = 6.2013$$

$$\sigma_4^2 = \frac{264{,}433.19}{44} - 6{,}003.0729 = 6.7723$$

$$p_{12} = \frac{\Sigma(X_1X_2)}{N} - c_1c_2$$

$$= \frac{63{,}198.42}{44} - 1{,}446.45396 = -10.1263$$

$$p_{13} = \frac{67{,}295.48}{44} - 1{,}541.0098 = -11.5671$$

$$p_{14} = \frac{66{,}550.84}{44} - 1{,}521.2403 = -8.7213$$

$$p_{23} = \frac{254{,}544.98}{44} - 5{,}782.1323 = 2.9808$$

$$p_{24} = \frac{251{,}246.54}{44} - 5{,}707.9535 = 2.1951$$

$$p_{34} = \frac{267{,}649.61}{44} - 6{,}081.0870 = 1.8586.$$

Substituting in the normal equations, we have:

$$-10.1263 = 9.2385b_{12.34} + 2.9808b_{13.24} + 2.1951b_{14.23}$$
$$-11.5671 = 2.9808b_{12.34} + 6.2013b_{13.24} + 1.8586b_{14.23}$$
$$-8.7213 = 2.1951b_{12.34} + 1.8586b_{13.24} + 6.7723b_{14.23}.$$

Solving these simultaneous equations[1] we secure the following values for the constants:

$$b_{12.34} = -0.460 \qquad b_{13.24} = -1.420 \qquad b_{14.23} = -0.749.$$

The required equation is, therefore,

$$x_1 = -0.460x_2 - 1.420x_3 - 0.749x_4.$$

This is the equation of regression of x_1 on x_2, x_3, and x_4. Any given values of the three independent variables (June temperature, July temperature, and August temperature) may be substituted in this equation, and the most probable value of the dependent variable (corn yield per acre) determined. In the equation as it stands, it should be noted, all the variables are expressed as deviations from their respective arithmetic means. For practical purposes it is advisable to have an equation in terms of the original values. In other words, it is desirable to shift the origin from the point of averages to the zero point on the original scales. This necessitates re-introducing the constant term a.

The value of a may be determined from the equation

$$A_1 = a + A_2b_{12.34} + A_3b_{13.24} + A_4b_{14.23}$$

where the A's represent the respective arithmetic means.[2] Inserting the proper values, we have

[1] Any method of solution may be employed. Perhaps the most convenient with three or more equations is the Doolittle method. This is explained in detail in Appendix A.

[2] This equation is derived from the first normal equation, as given on p. 537,

$$\Sigma(X_1) = Na + b_{12.34}\Sigma(X_2) + b_{13.24}\Sigma(X_3) + b_{14.23}\Sigma(X_4).$$

Replacing the absolute numbers X_1, X_2, etc., by their equivalents $x_1 + A_1$, $x_2 + A_2$, etc., we secure

$$\Sigma(x_1) + NA_1 = Na + b_{12.34}[\Sigma(x_2) + NA_2] + b_{13.24}[\Sigma(x_3) + NA_3] + b_{14.23}[\Sigma(x_4) + NA_4].$$

Since $\Sigma(x_1) = 0$, $\Sigma(x_2) = 0$, etc., these values disappear. Dividing through by N we obtain the equation presented above.

$$19.6341 = a + 73.6705(-0.46005) + 78.4864\,(-1.41967) + 77.4795(-0.74910).^{1}$$

Solving,

$$a = 222.99.$$

The equation of regression in terms of original values is, therefore,

$$X_1 = 222.99 - 0.460X_2 - 1.420X_3 - 0.749X_4.$$

COMPUTATION OF THE STANDARD ERROR OF ESTIMATE

Are estimates based upon this equation any more reliable than those based upon the equations previously derived, each of which referred to a single independent variable? To answer this question the value of the standard error must be computed. This will be represented in the present case by $S_{1.234}$, the subscripts referring to the single dependent variable (X_1) and the three independent variables. This value may be computed from the formula[2]

[1] The arbitrary origin is at zero on each of the original scales, hence $A_1 = c_1$, $A_2 = c_2$, etc. To ensure greater accuracy in solving for a, the values of the coefficients $b_{12.34}$, $b_{13.24}$, etc., are given to a greater number of decimal places than in the equation of regression.

[2] This formula may be derived as follows: Given an equation of the type

$$x_1 = b_{12.34}x_2 + b_{13.24}x_3 + b_{14.23}x_4$$

(in which the variables refer to deviations from the means) each residual may be computed from the equation

$$d = b_{12.34}x_2 + b_{13.24}x_3 + b_{14.23}x_4 - x_1. \qquad (1)$$

Multiplying throughout by d, and adding, we have

$$\Sigma(d^2) = b_{12.34}\Sigma(dx_2) + b_{13.24}\Sigma(dx_3) + b_{14.23}\Sigma(dx_4) - \Sigma(dx_1)$$

but it follows from the method of fitting that

$$\Sigma(dx_2) = 0$$
$$\Sigma(dx_3) = 0$$
$$\Sigma(dx_4) = 0$$

and, therefore, $\Sigma(d^2) = -\Sigma(dx_1)$. (2)

Multiplying each residual equation (1) by x_1 and adding, we have

$$\Sigma(dx_1) = b_{12.34}\Sigma(x_1x_2) + b_{13.24}\Sigma(x_1x_3) + b_{14.23}\Sigma(x_1x_4) - \Sigma(x_1^2).$$

Substituting the equivalent of $\Sigma(dx_1)$ in equation (2) we secure

(*Footnote continued on page 542.*)

$$S^2_{1.234} = \sigma_1^2 - b_{12.34}p_{12} - b_{13.24}p_{13} - b_{14.23}p_{14}.$$

Substituting the proper values, we have

$$\begin{aligned} S^2_{1.234} &= 44.6878 - 4.6586 - 16.4215 - 6.5331 \\ &= 17.0746 \\ S_{1.234} &= 4.13.^1 \end{aligned}$$

This is to be interpreted just as the standard error of estimate was interpreted in previous cases. The reliability of estimates based upon the mean value of X_1 is measured by σ_1, which has a value of 6.68. The reliability of estimates based upon the equation of net regression, when yield is considered as a function of temperature in June, July, and August, is measured by $S_{1.234}$ which has a value of 4.13. It is clear that estimates made from the equation are distinctly more reliable than those based upon a knowledge of X_1 alone. We have by no means accounted for all the factors that are responsible for variability in corn yield, but we have measured and reduced to precise terms the effect of three factors upon the yield of corn per acre in Kansas.

(*Footnote 2 continued from page 541.*)

$$\Sigma(d^2) = \Sigma(x_1^2) - b_{12.34}\Sigma(x_1x_2) - b_{13.24}\Sigma(x_1x_3) - b_{14.23}\Sigma(x_1x_4)$$

$$S^2_{1.234} = \frac{\Sigma(d^2)}{N} = \frac{\Sigma(x_1^2)}{N} - b_{12.34}\frac{\Sigma(x_1x_2)}{N} - b_{13.24}\frac{\Sigma(x_1x_3)}{N} - b_{14.23}\frac{\Sigma(x_1x_4)}{N}.$$

Since the variables refer to deviations from the means, we have

$$S^2_{1.234} = \sigma_1^2 - b_{12.34}p_{12} - b_{13.24}p_{13} - b_{14.23}p_{14}.$$

See Appendix A for a general derivation of these relations.

[1] For precise work, when the sample is small, allowance should be made in computing S for the number of constants in the equation of regression. Since there are four constants in the present equation, the 44 observations have but 40 degrees of freedom to deviate from the computed values. Denoting by $\bar{S}$ the corrected value of the standard error of estimate, and by m the number of constants in the equation of regression, Ezekiel gives

$$\bar{S}^2 = S^2\left(\frac{N}{N-m}\right)$$

applying this correction to the present measurements, we have

$$\begin{aligned} \bar{S}^2_{1.234} &= 17.0746\left(\frac{44}{44-4}\right) \\ &= 18.782 \\ \bar{S}_{1.234} &= 4.33. \end{aligned}$$

THE COEFFICIENT OF MULTIPLE CORRELATION

We have need now of our third measure, the abstract coefficient of correlation. The value of this coefficient, as we have seen, depends upon the relation between S and σ. It may be computed in the present instance from the formula

$$R^2_{1.234} = 1 - \frac{S^2_{1.234}}{{\sigma_1}^2}.$$

When the relationship between a single dependent variable and several independent variables is being studied, this measure is termed the coefficient of multiple correlation and is represented by the symbol R. The subscript to the left of the period relates to the dependent variable, while those to the right relate to the independent variables. Substituting in this formula the equivalent of $S^2_{1.234}$, we have

$$R^2_{1.234} = 1 - \frac{{\sigma_1}^2 - b_{12.34}p_{12} - b_{13.24}p_{13} - b_{14.23}p_{14}}{{\sigma_1}^2}$$

which reduces to[1]

$$R^2_{1.234} = \frac{b_{12.34}p_{12} + b_{13.24}p_{13} + b_{14.23}p_{14}}{{\sigma_1}^2}.$$

Inserting the proper values we have

$$R^2_{1.234} = \frac{4.6586 + 16.4215 + 6.5331}{44.6878}$$

$$R^2_{1.234} = .6179$$

$$R_{1.234} = .786.$$

For the same reason that estimates of ρ computed from samples must be corrected by making allowance for the number of constants in the regression equation, correction

[1] The coefficient of multiple correlation may also be derived from the general formula, which refers to an origin at zero on the original scales. This general formula is

$$R^2_{1.23\ldots n} = \frac{a\Sigma(X_1) + b_{12.34\ldots n}\Sigma(X_1X_2) + b_{13.24\ldots n}\Sigma(X_1X_3) + b_{14.23\ldots n}\Sigma(X_1X_4) + \ldots - Nc_1^2}{\Sigma(X_1)^2 - Nc_1^2}.$$

must be made in R. For if the number of constants is equal to the number of observations, R will necessarily equal 1. Using $\overline{R}$ to denote the corrected coefficient of multiple correlation and m to denote the number of constants in the equation of regression, Ezekiel gives

$$\overline{R^2} = 1 - \left\{(1 - R^2)\left(\frac{N-1}{N-m}\right)\right\}.$$

In the present example

$$\overline{R^2} = 1 - \left\{(1 - .6179)\left(\frac{44-1}{44-4}\right)\right\}$$
$$= .5892$$
$$\overline{R} = .768.$$

In later references to this illustration the uncorrected measure is used, though it is to be understood that the corrected measure provides a somewhat closer approximation to the true R than does the uncorrected coefficient.

The coefficient of multiple correlation is an index of the degree of relationship between a single dependent variable and a number of independent variables, in combination. It measures the degree to which variations in the dependent variable are related to the combined action of the other factors. Its significance may be clearer if all the independent variables are looked upon as constituting a single independent series. The coefficient is then seen to be a measure of the relationship between the dependent variable and the independent series, which is precisely what the coefficient of correlation is in the simpler case of two variables. In the multiple case the independent series has several component elements, but this fact does not alter the essential significance of the coefficient. No positive or negative sign is attached to R, it should be noted. In the present instance all of the independent variables are negatively correlated with corn yield, and a negative sign might be attached. The correlation could be positive, however, for some of

the independent variables, and negative for others. Because of this fact, R is always given without sign. The signs of the constants in the equation of net regression show which of the independent variables are positively correlated and which are negatively correlated with the dependent variable.

The sampling error of the coefficient of multiple correlation may be estimated from the formula

$$\sigma_R = \frac{1 - R^2}{\sqrt{N - m}}$$

where m is the number of constants in the equation of regression. A more accurate test of the significance of R may be applied with reference to Fisher's z-table, discussed in Chapter XV. The deviations of actual from computed values serve as a yardstick for testing the variability in X_1 that is attributable to X_2, X_3, and X_4, as the relationship is defined by the equation of regression. In common with other correlation problems, this one reduces to a comparison of variances.

The sum of the squares of the deviations of the observed values of X_1 from the computed values is 751.2824. The sum of the squares of the deviations of the computed values of X_1 from the mean value of X_1 is 1,214.9808. Since there are 44 observations, and since the equation of regression contains four constants, there are 40 degrees of freedom in the deviations from the regression function. The three coefficients of regression (other than the constant a) give three degrees of freedom to variation among the computed values of X_1. The test takes the following form.

Nature of variability	*Degrees of freedom*	*Sum of squared deviations*	*Mean square* σ^2	$Log_e\ \sigma^2$
Variation among computed values	3	1,214.9808	404.9936	6.0039
Deviation of observed from computed values	40	751.2824	18.7821	2.9329
	43	1,966.2632	Difference =	3.0710
			z =	1.5355

For $n_1 = 3$, $n_2 = 40$, the 1 per cent value of z, as derived from Appendix Table VI, is .7308. The present value is greatly in excess of this. The variation in X_1 attributable to the influence of X_2, X_3, and X_4 is clearly greater than the residual variability here used as the yardstick. The measure of correlation, R, is unquestionably significant.

COMPARISON OF MEASURES OF RELATIONSHIP

The degree to which our knowledge of the causes of variation in corn yield has been improved and the reliability of our estimates increased by taking account of the various factors in combination may be more readily appreciated if we bring together the various measures secured in the course of this analysis.

TABLE 129

A Comparison of Certain Measures Pertaining to the Corn Yield in Kansas

Basis of estimate	*Measure of reliability of estimate*	*Coefficient of correlation*
Arithmetic mean of $X_1 = 19.63$	$\sigma_1 = 6.68$	
$X_1 = 100\ 35 - 1.096X_2$	$S_{12} = 5.80$	$r_{12} = -.4984$
$X_1 = 166.07 - 1.866X_3$	$S_{13} = 4.81$	$r_{13} = -.6948$
$X_1 = 119.45 - 1.288X_4$	$S_{14} = 5.78$	$r_{14} = -.5013$
$X_1 = 222.99 - 0.460X_2 - 1.420X_3 - 0.749X_4$	$S_{1.234} = 4.13$	$R_{1.234} = .7861$

The value of S might be further reduced and the value of R correspondingly increased by bringing into the analysis other factors, such as rainfall during the growing months. The method which has been explained may be extended to cover any number of variables, one equation being added to the set of simultaneous equations for each additional variable introduced.

THE METHOD OF MULTIPLE CORRELATION VALID FOR LINEAR RELATIONSHIPS

One important condition has not been emphasized in the course of the preceding discussion. The validity of this method of multiple correlation depends upon the existence of a linear relationship between each pair of variables. Thus with four variables there were six pairings possible (i.e., six mean products were computed). If there had been a material departure from linearity in any of these six relationships the significance of the results would have been decreased. There would be no fallacy involved in the use of the equation under these conditions, but it would not furnish as good a basis for estimates as one which took account of the true relationship. In such a case the values of S and R would indicate that the estimates based upon the assumption of linear relationship were not very reliable.[1]

AN APPLICATION OF THE METHOD

Let us illustrate the use of the estimating equation. In the year 1933 the average June temperature in Kansas was 80.5° F., the average July temperature was 81.4° F., and the average August temperature was 76.8° F. What was the probable corn yield per acre? Substituting these values for X_2, X_3, and X_4 in the equation,

$$X_1 = 222.99 - 0.460X_2 - 1.420X_3 - 0.749X_4$$

we have

$$X_1 = 222.99 - (0.460 \times 80.5) - (1.420 \times 81.4) - (0.749 \times 76.8)$$

$$X_1 = 12.85.$$

The estimated yield for 1933 is thus 12.85 bushels per acre.

[1] An approach to problems of multiple correlation when the relationships among the variables are non-linear is explained by M. J. B. Ezekiel in the *Journal of the American Statistical Association*, Vol. XIX, N. S. No. 148, 1924, and in his book *Methods of Correlation Analysis*.

What are the limits within which we may expect the actual yield to fall, with respect to this estimate? The value of $S_{1.234}$ is 4.13 bushels. This means that the odds are 68 out of 100 that the actual yield will be within the limits 8.72 bushels (i.e., 12.85 − 4.13) and 16.98 bushels (i.e., 12.85 + 4.13). The actual yield in 1933 was 11.5 bushels per acre.

In this illustration we have used one of the years included in the study. The same method would be employed in making an estimate for a future year. (Additional elements of uncertainty are introduced, of course, whenever results secured for one period are applied to another time period.) Thus, from the temperatures in 1936 (76.7° in June, 85.5° in July, and 84.4° in August), an estimate of 3.1 bushels per acre is yielded by the regression equation employed above. This was a summer of exceptional heat and drought. The actual yield was 4.0 bushels per acre.

The Meaning of Partial or Net Correlation

In the preceding section we have sought to determine the degree to which corn yield in Kansas is affected by the temperature in June, July, and August, treating the three independent variables in combination. Our aim has been to measure their combined effect upon corn yield. There is a related problem, which in many studies may be of major importance. This is the determination of the relationship between a dependent variable and a single independent variable *when all other factors included in the study are held constant.* Concretely, what would be the effect upon corn yield of variations in July temperature, if June temperature and August temperature could be held constant? This is the problem of *net* or *partial correlation.*

It is obvious that if a method could be developed by which two variables could be isolated for separate study, it would add immeasurably to the analytical powers of the economist, and of social scientists in general. It would give

to the student in these fields that power to eliminate irrelevant influences and to concentrate his attention upon a single factor which is possessed by the chemist, for example. In studying the effect of one element upon another the chemist seeks to eliminate all other elements, and the effectiveness of his analysis depends in large part upon the degree to which it is possible thus to isolate the object of immediate interest.

It is not generally possible in economic analysis to eliminate all but one of the factors responsible for variations in a given series. The direct and indirect causes of a given economic phenomenon are too numerous and too complicated in their interaction for the economist ever to hope to emulate the chemist in reducing his problem to terms of but two variables. But, within certain limits, the statistician is able to employ the method of the physical scientist in holding constant certain factors while the effects of variations in another are studied. The methods which make this possible are among the most powerful of the instruments which the student of the social sciences possesses.

The method of partial correlation may be explained with reference to the problem of corn yield in Kansas. Our object is to determine the net correlation between corn yield and the temperature in each of the three months for which the average temperature is given.

DISTINCTION BETWEEN PARTIAL AND SIMPLE CORRELATION

It is important to distinguish between this problem and that faced in the ordinary measurement of relationship between two variables. We have already secured, as a description of the average relationship between corn yield and July temperature, the equation

$$X_1 = 166.07 - 1.866X_3$$

with

$$S_{13} = 4.81$$

and

$$r_{13} = -.6948.$$

These measures describe the relationship in question when all other factors are ignored. They are not taken account of. They are merely neglected. It is as though the chemist, in studying the reaction of one element to another, used a test tube containing various impurities, which he made no attempt to remove. The economist cannot, in general, locate and remove all the "impurities" in his problem, but he should recognize that his measures relate to such uncorrected data.

The Method of Partial Correlation

In seeking to determine the net correlation between corn yield and July temperature we attempt to secure a measure of the correlation which would prevail if other factors might be held constant. We shall take full account of the other factors we have studied, but we shall try to secure a measure influenced only by fluctuations in July temperature, in relation to corn yield.

One possible method of accomplishing this end may be suggested. If one possessed data covering a very long period we might be able to pick out a number of years during which the average temperatures in June and August remained unchanged. Let us say that we could find thirty years in all, during each of which the June temperature averaged 74° and the August temperature 78°. Corn yield and July temperature varied during these years. The relationship between July temperature and corn yield might now be measured, and it would be certain that the results would not be affected by the presence of fluctuations in June temperature and August temperature. Unfortunately, this method of holding certain factors constant cannot be employed. The data are too limited and too varied, in general, to enable us to pick from among them such figures as are appropriate to our purpose. Other methods of arriving at the same end are available, however.

As a first step, let us derive the equation defining the

relationship between corn yield as dependent variable and June temperature and August temperature as independent variables. This will be of the form

$$X_1 = a + b_{12.4}X_2 + b_{14.2}X_4.$$

We solve for the constants exactly as in the preceding example, except that variables X_1, X_2, and X_4 only are employed. The desired equation is

$$X_1 = 160.97 - 0.856X_2 - 1.010X_4.$$

We may determine the value of the standard error of estimate from the relation

$$S^2_{1.24} = \sigma_1^2 - b_{12.4}p_{12} - b_{14.2}p_{14}.$$

We secure

$$S^2_{1.24} = 27.2112$$
$$S_{1.24} = 5.22.$$

If corn yield per acre is estimated from June temperature and August temperature the standard error of estimate, or the standard deviation of the remaining variability, is 5.22 bushels. But we know that if corn yield is estimated from June, July, and August temperature, the standard error of estimate, or the standard deviation of the remaining variability, is 4.13 bushels. The measure of remaining or "unexplained" variability is reduced from 5.22 to 4.13 by the addition of July temperature (X_3) to the estimating equation, after account has already been taken of the influence of June temperature (X_2) and August temperature (X_4). The difference between these two measures may be taken to represent a relationship between X_1 and X_3 which is not affected by variations in X_2 and X_4.

We have seen that the degree of correlation between a dependent variable (X_1) and an independent variable (X_3) may be defined by the relation

$$r^2_{13} = 1 - \frac{S^2_{13}}{\sigma_1^2}.$$

Here the denominator of the fraction in the right-hand member defines the original variability of X_1, while the numerator of that fraction defines the variability of X_1 after account has been taken of the influence of X_3. In the present problem we have

$$r^2_{13} = 1 - \frac{23.1134}{44.6878} = .4828$$
$$r_{13} = -.695.$$

The coefficient of correlation is given the sign of b_{13}, the coefficient of regression.

In exactly the same way, we may say that the *net correlation* between X_1 and X_3, when the relationship is not affected by fluctuations in X_2 and X_4, is defined by the relation

$$r^2_{13.24} = 1 - \frac{S^2_{1.234}}{S^2_{1.24}}.$$

Here the denominator of the fraction in the right-hand member defines the variability remaining in X_1 after account has been taken of the influence of X_2 and X_4, while the numerator defines the variability remaining in X_1 after account has been taken of the influence of X_2, X_3, and X_4. *Numerator and denominator differ only because of the presence of correlation between X_1 and X_3 that is incremental to any correlation that may exist between X_1 on the one hand and X_2 and X_4 on the other.* If the equation

$$X_1 = 222.99 - 0.460X_2 - 1.420X_3 - 0.749X_4$$

gives estimates no more reliable than those derived from the equation

$$X_1 = 160.97 - 0.856X_2 - 1.010X_4$$

then numerator ($S^2_{1.234}$) and denominator ($S^2_{1.24}$) of the above fraction will be equal, and $r^2_{13.24}$ will have a value of zero. But if the equation containing X_2, X_3, and X_4 as independent variables gives better estimates than does the equation containing only X_2 and X_4, the numerator will be smaller

than the denominator, and $r^2_{13.24}$ will have a value other than zero. If the estimates based upon the three independent variables are in exact agreement with observed yields, $S^2_{1.234}$ will be equal to zero, and $r^2_{12.34}$ will have a value of unity.

Employing the values derived above, we have

$$r^2_{13.24} = 1 - \frac{17.0746}{27.2112} = .3725$$

$$r_{13.24} = -.610.$$

The coefficient of net correlation, $r_{13.24}$, is negative, having the same sign as the coefficient of net regression, $b_{13.24}$.

The quantity $r_{13.24}$ measures the degree of *correlation* between X_1 and X_3 when neither one is affected by variations in X_2 and X_4. It may be thought of, equally well, as a measure of the degree to which errors in estimating X_1 are reduced when use is made of X_3, after full account has already been taken of the influence of X_2 and X_4 on X_1.

The meaning of the symbols employed in the above demonstration should be clear from the context. As in the coefficients of net regression, the first of the subscripts to the left of the point (the primary subscripts) refers to the dependent variable; the second of the primary subscripts refers to the single independent variable to which the measure of net correlation applies specifically. The subscripts to the right of the point (the secondary subscripts) indicate the variables which are held constant for the purpose of the particular comparison being made. The number held constant is two in the present case, though it might be one, or any other number. Thus the general formula for the coefficient of net correlation between variables X_1 and X_3 would be

$$r^2_{13.2456\ \ldots\ n} = 1 - \frac{S^2_{1.23456\ \ldots\ n}}{S^2_{1.2456\ \ldots\ n}}.$$

The variable that is present in the numerator and absent in the denominator is the particular independent variable

that is being paired with the dependent variable for the purpose of measuring net relationship.

The coefficients of net correlation between X_1 and each of the other independent variables may be derived in similar fashion. Thus

$$r^2_{12.34} = 1 - \frac{S^2_{1.234}}{S^2_{1.34}}$$

$$r^2_{14.23} = 1 - \frac{S^2_{1.234}}{S^2_{1.23}}.$$

In each case the difference between numerator and denominator of the fraction in the right-hand member measures the net reduction in the variability of X_1 which is associated with a relationship between X_1 and a single independent variable, account having already been taken of the influence of two other variables.

It is clear that such measurements as these are *net* only with respect to the variables represented by the secondary subscripts. The coefficient $r_{12.34}$ measures the degree of relation between X_1 and X_2 when X_3 and X_4 are held constant. There may be many other factors affecting X_1 and X_2; the disturbing influences of such factors have not been eliminated. These other factors still muddy the water of analysis. Ignoring them is not the same as holding them constant. Only by direct measurement and inclusion in the study, as was done with X_3 and X_4, may the influence of additional variables be effectively eliminated.

Another Method of Computing Coefficients of Partial Correlation

Obviously a whole series of coefficients of net correlation may be computed in dealing with a number of variables. In deriving a number of such measurements a method may be utilized which differs somewhat from that employed above, and which has certain advantages in the way of systematic arrangement.

A simple coefficient of correlation relating to but two variables is termed a *coefficient of zero order*. Such coefficients are represented by symbols of the type r_{12}, r_{24}, etc. Coefficients of net correlation which relate to two variables, while a single additional variable is held constant, are termed *coefficients of the first order*, and are represented by symbols such as $r_{12.3}$, $r_{24.3}$, etc. Similarly, we may have coefficients of the second, third, fourth, or nth order, depending upon the number of variables held constant while the relationship between a single dependent and a single independent variable is being measured.

It is possible to derive each coefficient of partial correlation from those of the next lower order. Thus a coefficient of the first order may be derived from the relation

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{(1 - r^2_{13})^{\frac{1}{2}} (1 - r^2_{23})^{\frac{1}{2}}}.$$

For a coefficient of the second order

$$r_{12.34} = \frac{r_{12.3} - r_{14.3} \cdot r_{24.3}}{(1 - r^2_{14.3})^{\frac{1}{2}} (1 - r^2_{24.3})^{\frac{1}{2}}}.$$

As a general equation for a coefficient of net correlation of any order,[1] we have

$$r_{12.345\ldots n} = \frac{r_{12.345\ldots(n-1)} - r_{1n.345\ldots(n-1)} \cdot r_{2n.345\ldots(n-1)}}{(1 - r^2_{1n.345\ldots(n-1)})^{\frac{1}{2}} (1 - r^2_{2n.345\ldots(n-1)})^{\frac{1}{2}}}.$$

Thus it is possible, starting with the zero order coefficients of correlation, to compute all higher order coefficients successively. The mere arithmetic of calculation would be laborious, but certain prepared tables reduce these computa-

[1] It will be noted that in an equation used in computing a coefficient of partial correlation the three r's in the numerator of the right-hand member have the same secondary subscripts, and that these secondary subscripts are one less in number than the secondary subscripts of the left-hand member; that the first r in the numerator has the same primary subscripts as the left hand member; that the second and third r's in the numerator have primary subscripts composed of one of the primary subscripts of the left-hand member plus the missing secondary subscript; that the two r's in the denominator are the same as the second and third r's in the numerator.

tions to a minimum.[1] The method may be illustrated, using the data of the preceding problem.

In the present case we require three coefficients of the second order, $r_{12.34}$, $r_{13.24}$, and $r_{14.23}$. These will serve as measures of the net correlation between corn yield and temperature in each of the three critical months. The formula from which the first of these measures may be computed was given above. For the second, we have

$$r_{13.24} = \frac{r_{13.2} - r_{14.2} \cdot r_{34.2}}{(1 - r^2_{14.2})^{\frac{1}{2}} (1 - r^2_{34.2})^{\frac{1}{2}}}$$

and for the third

$$r_{14.23} = \frac{r_{14.2} - r_{13.2} \cdot r_{43.2}}{(1 - r^2_{13.2})^{\frac{1}{2}} (1 - r^2_{43.2})^{\frac{1}{2}}}.$$

But each of these values may be derived from a slightly different grouping of first order coefficients. We may use the three formulas

$$r_{12.34} = \frac{r_{12.4} - r_{13.4} \cdot r_{23.4}}{(1 - r^2_{13.4})^{\frac{1}{2}} (1 - r^2_{23.4})^{\frac{1}{2}}}$$

$$r_{13.24} = \frac{r_{13.4} - r_{12.4} \cdot r_{32.4}}{(1 - r^2_{12.4})^{\frac{1}{2}} (1 - r^2_{32.4})^{\frac{1}{2}}}$$

$$r_{14.23} = \frac{r_{14.3} - r_{12.3} \cdot r_{42.3}}{(1 - r^2_{12.3})^{\frac{1}{2}} (1 - r^2_{42.3})^{\frac{1}{2}}}.$$

By employing both methods in computing each second order coefficient a check upon the calculations is afforded.

COMPUTATION OF FIRST ORDER COEFFICIENTS

The second order coefficients cannot be computed until all necessary first order coefficients have been secured. The necessary equations, of the type

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{(1 - r^2_{13})^{\frac{1}{2}} (1 - r^2_{23})^{\frac{1}{2}}},$$

may be constructed from the general formula for coefficients of partial correlation. Since several of these values must be computed, a systematic arrangement should be employed.

[1] J. R. Miner, *Tables of* $\sqrt{1 - r^2}$ *and* $1 - r^2$ *for use in Partial Correlation and in Trigonometry*, Johns Hopkins Press, Baltimore, Md., 1922.

TABLE 130

Illustrating the Computation of First Order Coefficients of Partial Correlation

(Kansas corn yield and temperature)

r 0 *Order*		$(1-r^2)^{\frac{1}{2}}$	*Product term of numerator*	*Whole numerator*	*Denominator*	*r* 1*st Order*	
Subscript	*Coefficient*					*Subscript*	*Coefficient*
12	− .4984		− .2736	− .2248	.6611	12.3	− .3400
13	− .6948	.7192					
23	+ .3938	.9192					
14	− .5013		− .1993	− .3020	.6890	14.3	− .4383
13	− .6948	.7192					
43	+ .2868	.9580					
24	+ .2775		+ .1129	+ .1646	.8806	24.3	+ .1869
23	+ .3938	.9192					
43	+ .2868	.9580					
13	− .6948		− .1963	− .4985	.7969	13.2	− .6255
12	− .4984	.8669					
32	+ .3938	.9192					
14	− .5013		− .1383	− .3630	.8329	14.2	− .4358
12	− .4984	.8670					
42	+ .2775	.9607					
34	+ .2868		+ .1093	+ .1775	.8831	34.2	+ .2010
32	+ .3938	.9192					
42	+ .2775	.9607					
12	− .4984		− .1391	− .3593	.8313	12.4	− .4322
14	− .5013	.8653					
24	+ .2775	.9607					
13	− .6948		− .1438	− .5510	.8290	13.4	− .6647
14	− .5013	.8653					
34	+ .2868	.9580					
23	+ .3938		+ .0796	+ .3142	.9204	23.4	+ .3414
24	+ .2775	.9607					
34	+ .2868	.9580					

The procedure in computing each first order coefficient is simple. Three zero order coefficients are necessary for each calculation. These should be arranged in the table in the order in which they occur in the numerator of the fraction from which the required coefficient is to be computed. The numerator of this fraction is secured by subtracting from the first zero order coefficient the product of the other two. This product term appears in one column of the table. The denominator of the fraction is the product of two terms of the type $\sqrt{1 - r^2}$, derived from the second and third coefficients in each group of three. The tabular arrangement of Table 130 on page 557 permits these computations to be carried forward systematically.

The coefficient $r_{23.4}$ is, of course, identical with $r_{32.4}$; $r_{34.2}$ is identical with $r_{43.2}$, etc. It is unnecessary to duplicate the work of computation with respect to these measures.

COMPUTATION OF SECOND ORDER COEFFICIENTS

From these first order coefficients the three required second order coefficients may be secured by methods analogous to those employed above. The computations are shown in Table 131. As a check upon the calculations each required measure is computed from two different combinations of the first order coefficients.

The value of $r_{13.24}$, it will be noted, is the same as that derived from the relation between $S_{1.24}$ and $S_{1.234}$.

The meaning of such coefficients as these was explained in the earlier section dealing with this problem. The following summary of results reveals the gain in knowledge which has resulted from the above analysis.

$r_{12} = -.4984$	$r_{12.34} = -.2923$
$r_{13} = -.6948$	$r_{13.24} = -.6101$
$r_{14} = -.5013$	$r_{14.23} = -.4057$

It is clear that the net effect of June temperature upon corn yield is distinctly less than was indicated by the simple

TABLE 131

Illustrating the Computation of Second Order Coefficients of Partial Correlation

(Kansas corn yield and temperature)

r 1st Order		$(1-r^2)^{\frac{1}{2}}$	*Product term of numerator*	*Whole numerator*	*Denominator*	*r 2nd Order*	
Subscript	*Coefficient*					*Subscript*	*Coefficient*
12.3	− .3400		− .0819	− .2581	.8830	12.34	− .2923
14.3	− .4383	.8988					
24.3	+ .1869	.9824					
13.2	− .6255		− .0876	− .5379	.8816	13.24	− .6101
14.2	− .4358	.9000					
34.2	+ .2010	.9796					
14.2	− .4358		− .1257	− .3101	.7643	14.23	− .4057
13.2	− .6255	.7802					
43.2	+ .2010	.9796					
12.4	− .4322		− .2269	− .2053	.7022	12.34	− .2924
13.4	− .6647	.7471					
23.4	+ .3414	.9399					
13.4	− .6647		− .1476	− .5171	.8476	13.24	− .6101
12.4	− .4322	.9018					
32.4	+ .3414	.9399					
14.3	− .4383		− .0635	− .3748	.9238	14.23	− .4057
12.3	− .3400	.9404					
42.3	+ .1869	.9824					

correlation. This is so because there is a positive correlation between temperature in June and temperature in July and August, so that the crude correlation of two variables alone shows June temperature as more important than it really is. For the same reason, all the net coefficients are less than the simple coefficients, though it is still apparent that July temperature is far more important, in relation to corn yield, than the temperature in either of the other months.

The coefficients of net correlation are net, of course, only with respect to the variables actually taken account of, and held constant. Thus there may be other factors, such as rainfall in June, July, or August, which affect corn yield and which are correlated with the temperature during these months. Were these included the various coefficients of net correlation might have different values from those given.

The sampling error of a coefficient of partial correlation may be estimated from the same general relations that hold for zero order coefficients, except that the factor $N - 1$ must be further reduced by the number of variables held constant. Thus for $r_{12.34}$ we have

$$\sigma_{r_{12.34}} = \frac{1 - r^2_{12.34}}{\sqrt{N - 3}}.$$

A MEASURE OF VARIABILITY

Having these coefficients of net correlation, another measure of some importance may be computed. This is a measure of the variability of a single character while a number of related variables are held constant. Thus the question might arise: If we could hold constant the temperature in Kansas in June, July, and August, what would be the variability of the corn yield? In other words: If we could eliminate such variability in corn yield as is due to variability in temperature, what fluctuations would remain in the yield of corn? This measure of variability is represented by the symbol $\sigma_{1.234\ldots n}$. It is termed the standard deviation of order n.

This measure may be computed from the general equation

$$\sigma^2_{1.23\ldots n} = \sigma_1{}^2(1 - r^2_{12})(1 - r^2_{13.2})(1 - r^2_{14.23}) \ldots (1 - r^2_{1n.23\ldots n-1}).$$

Applying this formula to the results of the study of corn yield, we have

$$\sigma^2_{1.234} = 44.6878[1 - (-.4984)^2][1 - (-.6255)^2][1 - (-.4057)^2]$$
$$\sigma^2_{1.234} = 17.0797$$
$$\sigma_{1.234} = 4.13.$$

Referring back to the discussion of this problem we find that the values of $\sigma_{1.234}$ and $S_{1.234}$ are identical. That is, the standard deviation of variable X_1, when variables X_2, X_3, and X_4 are held constant, is merely the standard deviation of observed values from computed values of X_1. It is the standard error of estimate, when estimates are based upon the factors X_2, X_3, and X_4. The reason for this is obvious. The variability of the original series is reduced to the extent that estimates based upon the equation of relationship approximate the actual values. The variability which remains is due to differences between these estimates and the actual values. But these differences are merely the residual deviations, from which S is computed. A realization of the identity of these two measures may assist in making their meaning clear.

Since $\sigma_{1.234}$ and $S_{1.234}$ are identical, the coefficient of multiple correlation, $R_{1.234}$, may be computed from the equation

$$R^2_{1.23\ \ldots\ n} = 1 - \frac{\sigma^2_{1.23\ \ldots\ n}}{\sigma_1{}^2}$$

or, using the formula for $\sigma^2_{1.234\ \ldots\ n}$, from the equation

$$1 - R^2_{1.23\ \ldots\ n} = (1 - r^2_{12})(1 - r^2_{13.2})(1 - r^2_{14.23}) \ldots (1 - r^2_{1n.23\ \ldots\ (n-1)}).$$

BETA COEFFICIENTS

The several coefficients of regression in an equation of multiple regression are, in effect, weights applied to the different independent variables in estimating the successive values of the dependent variable. Usually these coefficients of regression are not comparable, because the independent factors are expressed in different units, or because they differ in variability. It is often desirable to reduce the

coefficients of regression to comparable terms. This may be done by expressing dependent and independent variables alike in units of their respective standard deviations. The coefficients of regression are then called *beta coefficients*, and are represented by the symbols $\beta_{12.34}$, $\beta_{13.24}$, etc.

In terms of a simple two-variable problem, we have

$$x_1 = b_{13}x_3.$$

If we change to standard deviation units we must divide both sides of the equation by σ_1 and by σ_3. This gives

$$\frac{x_1}{\sigma_1\sigma_3} = \frac{b_{13}}{\sigma_1}\left(\frac{x_3}{\sigma_3}\right)$$

or

$$\frac{x_1}{\sigma_1} = \left(b_{13}\frac{\sigma_3}{\sigma_1}\right)\left(\frac{x_3}{\sigma_3}\right).$$

The desired Beta coefficient is, then,

$$\beta_{13} = b_{13}\left(\frac{\sigma_3}{\sigma_1}\right).$$

For the corn yield example, we have

$$\beta_{13} = -1.866\left(\frac{2.49}{6.68}\right) = -.696.$$

This may be taken to mean that with an increase of one standard deviation in X_3 (July temperature), the yield of corn decreased .696 of one standard deviation.

These measurements are particularly useful in analyses involving more than two variables. Here the relationships between the beta coefficients and the coefficients of net regression are similar to those indicated for the two-variable problem. Thus

$$\beta_{12.34} = b_{12.34}\left(\frac{\sigma_2}{\sigma_1}\right)$$

$$\beta_{13.24} = b_{13.24}\left(\frac{\sigma_3}{\sigma_1}\right)$$

$$\beta_{14.23} = b_{14.23}\left(\frac{\sigma_4}{\sigma_1}\right).$$

Substituting the required values in these equations, we have

$$\beta_{12.34} = -.209$$
$$\beta_{13.24} = -.529$$
$$\beta_{14.23} = -.292.$$

The second of these coefficients may be taken to mean that with an increase of one standard deviation in July temperature, when June and August temperatures are held constant, corn yield decreases .529 of one standard deviation. The other coefficients have similar meanings.

The beta coefficients relate to factors expressed in comparable units and similar in respect of variability. A fluctuation of one standard deviation in X_2 may be taken to be equal to a fluctuation of one standard deviation in X_3. The coefficients defining the changes in X_1 that are likely to accompany these equal movements in X_2 and X_3 have obvious significance.

CERTAIN LIMITATIONS

The measures we have described in dealing with problems of multiple and partial correlation are valid on the assumption that the relationships among the different variables are in all cases linear. Thus with four variables six different pairs may be obtained. The regression in each of these six cases should be linear if combined or net effects are to be studied by the methods outlined above. If the regression is non-linear when natural numbers are dealt with, it may be possible to secure linear relationships by correlating logarithms or reciprocals. Thus we might derive an estimating equation of the type

$$\text{Log } X_1 = a + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$$

if the relation between X_1 in logarithmic form and each of the other variables in the original arithmetic form were linear. The corresponding measures, S and R, would then

relate to ratios, as in the examples given in the following chapter.[1]

One other important limitation should be noted. Coefficients of multiple or of net correlation based upon a large number of variables have little significance unless the number of observations be large. Misleadingly high values will be secured when studies involving many variables are based upon small samples. (Application of the corrections referred to in the text will prevent misinterpretation, in such cases.) Within the limits set by these restrictions, the methods of multiple and partial correlation constitute very powerful instruments of economic analysis.

REFERENCES

Bean, L. H., "A Simplified Method of Graphic Curvilinear Correlation," *Journal of the American Statistical Association*, Dec., 1929.

Bowley, A. L., *Elements of Statistics*, Part II, Chap. 8.

Camp, B. H., *The Mathematical Part of Elementary Statistics*, Part II, Chap. 6.

Davies, G. R. and Yoder, D., *Business Statistics*, Chap. 7.

Elderton, W. P., *Frequency Curves and Correlation*, Chap. 13.

Ezekiel, M. J. B., *Methods of Correlation Analysis*, Chaps. 12–16, 18–21.

Haas, G. C., "Sale Prices as a Basis for Farm Land Appraisal," *Technical Bulletin No. 9.* University of Minnesota Agricultural Experiment Station.

Kelley, T. L., *Statistical Method*, Chap. 11.

Kelley, T. L., *Partial and Multiple Correlation* (in *Handbook of Mathematical Statistics*, Rietz, H. L., ed., Chap. 9).

[1] Considerable use has been made in agricultural economics of a method of measuring curvilinear multiple correlation developed by Mordecai Ezekiel, and of a simplified graphic procedure devised by Louis H. Bean. These procedures provide flexible instruments of analysis particularly well adapted to exploratory work in the study of relations among variable quantities. The conclusions to which they may lead are limited by the fact that the use of free-hand methods of defining regression curves introduces subjective elements into the analytical work. For a discussion of these methods see the references to works by Mordecai Ezekiel and Louis H. Bean at the end of this chapter. A critical appraisal of the procedure is given in the article by Malenbaum and Black, which is there cited.

Malenbaum, Wilfred and Black, John D., "The Use of the Short-Cut Graphic Method of Multiple Correlation," *Quarterly Journal of Economics*, Nov., 1937.

Miner, J. R., *Tables of* $\sqrt{1 - r^2}$ *and* $1 - r^2$ *for Use in Partial Correlation and Trigonometry.*

Pearl, R., *Medical Biometry and Statistics*, Chap. 15.

Snedecor, G. W., *Statistical Methods*, Chap. 13.

Tippett, L. H. C., *The Methods of Statistics*, Chap. 11.

Tolley, H. R. and Ezekiel, M. J. B., "A Method of Handling Multiple Correlation Problems," *Journal of the American Statistical Association*, Dec., 1923.

Waugh, A. E., *Elements of Statistical Method*, Chap. 11.

Yule, G. U. and Kendall, M. G., *An Introduction to the Theory of Statistics*, Chap. 14.

CHAPTER XVII

THE MEASUREMENT OF RELATIONSHIP AND THE PROBLEM OF ESTIMATION

It is no great exaggeration to say that quantitative method in economics and business centers about the problem of estimation. Equations of regression, measures of standard error and coefficients of correlation are of interest largely because of their bearing upon the practical problems of determining probable production, probable price, probable business changes. It should not be understood from this that the problem of estimation relates only to attempts to forecast future changes. We make an estimate whenever we seek to determine the most probable value from a number of different observations, or whenever we employ an equation which describes the relation between two or more variables. The value of statistical technique rests in large part upon its practical utility in the making of estimates.

This object has been definitely to the fore in the preceding chapters, which dealt with methods by which the value of one variable might be estimated from a given value of another. We may, at this point, briefly summarize certain assumptions upon which the validity of this method rests.

Some Assumptions Involved in the Making of Estimates

In earlier chapters it has been pointed out that the most probable value of a series of observations is their arithmetic mean. Given a normal distribution about the mean, the standard deviation affords an exact measure of the proba-

bilities involved in basing estimates upon the mean. Similarly, the standard error of estimate affords an exact measure of the probabilities involved in basing estimates upon an equation of regression, again upon the assumption that the distribution about the line of regression is normal. The significance and usefulness of the equation of regression may be determined by comparing the standard error of estimate of a given variable with the standard deviation.

From the relation between these two values, moreover, an abstract measure of relationship, the coefficient or index of correlation, may be computed. This coefficient, or index, is a thoroughly valid and accurate measure only if the distribution about the line of regression and the distribution about the mean are normal, or approximately so. Pronounced departures from the normal type lessen the significance of these measures.

In the foregoing discussion we have been concerned with *arithmetic* values throughout. In speaking of estimates based upon the mean we referred to the arithmetic mean. The distributions about the mean and about the line of regression are assumed to be normal when deviations are measured arithmetically. The standard deviation and the standard error of estimate are in arithmetic terms, referring to absolute values. But may we assume that all the distributions we deal with in economic analysis are of the arithmetic type? Should estimates be made and errors of estimate measured only in arithmetic terms? If they should not be so limited, are the methods developed above capable of adaptation to other distributions? These questions may best be answered in terms of a specific problem.

A Problem of Estimation: Logarithmic and Ratio Values

In Table 132 the production and price of oats in the United States from 1881 to 1913 are recorded. Appropriate lines of trend were fitted to these series and the ratios

TABLE 132

Production and Price of Oats in the United States

Year	*Production of oats in U. S. (millions of bu.)*	*Straight line trend of production* [1]	*Ratio of actual production to trend value*	*Price of oats in Chicago (cents per bu.)*	*Straight line trend of price* [2]	*Ratio of actual price to trend value*
1881	416	448	.929	47	36.0	1.30
1882	488	471	1.036	37	35.3	1.05
1883	571	494	1.156	31	34.6	.90
1884	583	517	1.128	29	34.0	.85
1885	629	540	1.165	28	33.2	.84
1886	624	563	1.108	25	32.5	.77
1887	659	586	1.124	30	31.2	.96
1888	701	609	1.151	24	30.5	.79
1889	751	632	1.188	24	29.8	.81
1890	523	655	.798	43	29.0	1.48
1891	738	678	1.088	31	28.3	1.10
1892	661	701	.943	30	27.5	1.09
1893	639	724	.882	31	26.8	1.16
1894	662	747	.886	28	26.1	1.07
1895	824	770	1.070	19	25.3	.75
1896	780	793	.983	18	23.6	.76
1897	791	816	.969	24	25.0	.96
1898	843	839	1.005	25	26.4	.95
1899	926	862	1.074	23	27.8	.83
1900	914	885	1.033	25	29.2	.86
1901	778	908	.857	42	30.6	1 37
1902	1,053	931	1.131	33	32.0	1.03
1903	869	954	.911	38	33.4	1.14
1904	1,009	977	1.033	30	34.8	.86
1905	1,090	1,000	1.090	31	36.2	.86
1906	1,036	1,023	1.013	39	37.6	1.04
1907	805	1,046	.770	51	39.0	1.31
1908	851	1,069	.796	52	40.4	1.29
1909	1,068	1,092	.978	43	41.8	1.03
1910	1,186	1,115	1 064	35	43.2	.81
1911	922	1,138	.810	51	44.6	1.14
1912	1,418	1,161	1.221	37	46.0	.80
1913	1,122	1,184	.948	41	47.4	.87

[1] This line of trend was fitted to data covering a longer period than that included in the present study.

[2] The entire period has been broken into two parts, 1881 to 1895 and 1896 to 1913. A straight line of trend was fitted by H. B. Killough to the data of each period.

of the actual values of the items in each series to the trend values determined.

It is desired to measure the relation between these two variables. A hyperbolic curve of the general type $Y = aX^b$ appears to be an appropriate form to employ in describing such a relationship. To fit this curve by the method of least squares, the equation must be reduced to the logarithmic form

$$\log Y = \log a + b \log X.$$

The normal equations required in fitting a curve of this type, are

$$\text{I} \quad \Sigma(\log Y) = N \log a + b\,\Sigma(\log X)$$
$$\text{II} \quad \Sigma(\log X \cdot \log Y) = \log a \Sigma(\log X) + b\Sigma(\log^2 X).$$

The values necessary for the solution of these equations are determined from Table 133.[1]

From this table we have

$$N = 33$$
$$\Sigma(\log Y) = -.32849 \qquad \Sigma(\log X \cdot \log Y) = -.1143005$$
$$\Sigma(\log X) = .037535 \qquad \Sigma(\log^2 X) = .096423.$$

Substituting in the normal equations, we secure

$$-.32849 = 33 \log a + .037535b$$
$$-.1143005 = .037535 \log a + .096423b.$$

Solving

$$\log a = -.00861$$
$$b = -1.18206.$$

The required equation is

$$\log Y = (9.99139 - 10) - 1.18206 \log X$$

or

$$Y = .9804X^{-1.18206}.$$

[1] I am indebted to Prof. H. B. Killough of Brown University for permission to use the data presented in Tables 132 and 133. The figures are taken from his comprehensive study of the factors affecting oat prices.

This is the equation which describes the average relationship between the production and the price of oats (when the actual figures for each are expressed as ratios to the respective lines of trend). The corresponding curve is plotted in Fig. 88 on page 592.

TABLE 133

Computation of Values Required in Fitting a Curve to Data of Oat Production and Prices

EXAMPLE I

(1) Year	(2) Ratio of price to trend Y	(3) Ratio of production to trend X	(4) $\log Y$	(5) $\log X$	(6) $\log^2 Y$	(7) $\log^2 X$	(8) $\log Y \cdot \log X$
1881	1.30	929	.1139434	.9680157 − 1	.01298310	.001022995	− .0036444
1882	1.05	1 036	.0211893	.0153598	.00044899	.000235923	.0003255
1883	90	1 150	9542425 − 1	.0629578	.00209375	.003963685	− .0028808
1884	85	1 128	.9294189 − 1	0523091	.00498169	.002736242	− .0036920
1885	.84	1 165	.9242793 − 1	.0663259	.00573362	.004399125	− 0050222
1886	77	1 108	8864907 − 1	.0445398	.01288436	.001983794	− 0050557
1887	96	1 124	9822712 − 1	.0507663	00031431	.002577217	− .0000000
1888	.79	1 151	.8976271 − 1	.0610753	.01048021	.003730192	− 0062524
1889	81	1 188	.9084850 − 1	0748164	.00837500	.005597494	− .0068468
1890	1 48	798	1702617	.9020029 − 1	.02898905	.009603432	− .0166852
1891	1 10	1 088	.0413927	.0366289	.00171336	.001341676	.0015162
1892	1 09	.943	.0374265	9745117 − 1	.00140074	.000649653	− .0009539
1893	1 16	882	0644580	.9454686 − 1	.00415483	.002973674	− .0035150
1894	1 07	886	.0293838	.9474337 − 1	.00086341	.002763216	− .0015446
1895	75	1 070	8750613 − 1	.0293838	.01560968	.000863408	− 0036712
1896	.76	.933	8808136 − 1	.9925535 − 1	.01420540	.000055450	0008875
1897	96	969	.9822712 − 1	.9863238 − 1	.00031431	.000187038	.0002425
1898	95	1 005	9777236 − 1	.0021661	.00049624	.000004692	− 0000483
1899	83	1 074	.9190781 − 1	.0310043	.00654835	.000961267	− 0025089
1900	86	1.033	.9344985 − 1	.0141003	.00429045	.000198818	− 0009236
1901	1.37	857	1367206	.9329808 − 1	.01725316	.004491573	− .0091629
1902	1 03	1 131	.0128372	.0534626	.00016479	.002858250	0006863
1903	1 14	911	.0569049	.9595184 − 1	.00323817	.001638760	− 0023036
1904	.86	1 033	.9344985 − 1	.0141003	.00429045	.000198818	− 0009236
1905	86	1 090	.9344985 − 1	.0374265	.00429045	.001400743	− 0024515
1906	1 04	1 013	.0170333	.0056004	.00029013	.000031465	.0000955
1907	1 31	770	.1172713	.8864907 − 1	.01375256	.012884361	− .0133113
1908	1 29	.796	.1105897	.9009131 − 1	.01223008	.009818214	− .0109580
1909	1 03	978	.0128372	.9903389 − 1	.00016479	.000093337	− 0001240
1910	.81	1 064	.9084850 − 1	.0269416	.00837500	.000725850	− 0024656
1911	1 14	810	0569049	.9084850 − 1	00323817	.008374812	− 0052076
1912	.80	1.221	.9030900 − 1	.0867157	.00939155	.007519613	− 0084036
1913	87	.948	.9395193 − 1	.9768083 − 1	.00365792	.000537855	.0014027
Total	32 83	33.338	17.6715068 − 18	14.0375350 − 14	.21721807	.096422642	− .1194567
							+ .0051562
							− .1143005

THE STANDARD ERROR OF ESTIMATE IN LOGARITHMIC TERMS

How reliable is this equation? With what degree of confidence may estimates be based upon it? To answer these questions we must compute the standard error, S. Since the fitting process was carried through in terms of

logarithms, the standard error may be computed in the same terms. Following the procedure explained in earlier sections with reference to the straight line and the potential series, we may derive the following equation relating to the logarithmic curve just fitted:

$$S^2_{\log y} = \frac{\Sigma(\log^2 Y) - \log a\Sigma(\log Y) - b\Sigma(\log X \cdot \log Y)}{N}.$$

Substituting the proper values, we have

$$S^2_{\log y} = \frac{.21721807 - (-.00861 \times -.32849) - (-1.18206 \times -.1143005)}{33}$$

$$= \frac{.07927928}{33}$$

$$S^2_{\log y} = .0024024$$
$$S_{\log y} = .04901.$$

The standard error of estimate, in the form of a logarithm, is .04901. As long as we deal with logarithms, this is to be interpreted precisely as is the standard error with respect to other curves. Assuming a normal distribution of logarithms about the curve which describes the average relationship, the chances are 68 out of 100 that the logarithm of a given estimate will not differ from the logarithm of the actual value by more than .04901, 95 out of 100 that the logarithm of the given estimate will not differ from the logarithm of the actual value by more than .09802, and 99.7 out of 100 that the logarithm of the given estimate will not differ from the logarithm of the actual value by more than .14703.

INTERPRETATION OF THE STANDARD ERROR OF ESTIMATE; ZONES OF ESTIMATE

What does this mean in terms of actual values? It means, simply, that we are dealing throughout in terms of ratios instead of absolute figures. The difference between the logarithms of two numbers is the logarithm of the ratio

of one of the original numbers to the other. Thus the absolute value of S in a given case will depend upon the magnitude of the values with which we are dealing. If the user desires to reduce S to absolute values, it must be done always with reference to a given estimate. That is, a given value of X is substituted in the equation of average relationship and the corresponding value of Y estimated. If the logarithmic equation is used, this estimate will be in the form of a logarithm. To the logarithm of the estimate add the value of $S_{\log y}$. The anti-logarithm of the number thus secured will give the upper limit of a zone extending a distance equal to S above the line of regression. From the logarithm of the estimate subtract the value of $S_{\log y}$. The anti-logarithm of the number thus secured will give the lower limit of a zone extending a distance equal to S below the line of regression. The odds are 68 out of 100 that the value of Y in the given case will fall within the limits thus marked out. The absolute limits corresponding to $2S$ and $3S$ may be similarly determined.

The zone thus marked out with respect to a logarithmic curve will differ materially from the similar zones already described in dealing with simple linear equations. In the simple case a zone extending $1S$ on each side of the estimating curve has the same absolute width throughout its length, and is centered always at the line of regression. The logarithmic zone, when measured in natural numbers, is of varying width, and, moreover, is not of the same width on each side of the plotted curve. It is true, however, that the ratios on the two sides of the curve are always equal. That is, the ratio of a value $1S$ less than the computed value to the computed value is the same as the ratio of the latter to a value $1S$ greater. And when the curves are plotted on paper ruled logarithmically, the zone included within a distance $1S$ on each side of the plotted curve takes the symmetrical form found in the earlier and simpler cases. A person accustomed to thinking in terms of ratios

and to the use of logarithmic paper can readily interpret this measure.

THE STANDARD ERROR OF ESTIMATE IN TERMS OF RATIOS

Since the ratios are equal throughout, the standard error of estimate may be expressed in ratio terms. In the present example we have

$$S_r = \text{anti-log } S_{\log y} = \text{anti-log } .04901 = 1.12$$

where S_r is used to represent the standard error of estimate in terms of ratios. $S_{\log y}$, as derived above, is positive, hence the ratio exceeds unity. It is the ratio of the larger number to the smaller. What does it mean? It means that in 68 cases out of 100 the actual value, if it exceed the estimate, will not exceed it by more than 12 per cent, and, if it fall below the estimate, will stay within a limit such that the estimate will not be more than 12 per cent greater than the actual value. This is not a convenient form, since this ratio always expresses the larger value in terms of the smaller value. It would be more convenient to have it always in terms of a percentage of the estimate. This may be done by putting $S_{\log y}$ in negative terms, and getting the corresponding natural value. The value $-.04901 = 9.95099 - 10$, which is the logarithm of .8933. In this form the ratio is based upon the relation of the smaller to the larger number. To make S_r readily intelligible we may combine the two, writing

$$S_r = .89 \text{ to } 1.12.$$

Interpreting this, it means that, given a normal distribution, in 68 cases out of 100 the actual value will not be less than 89 per cent of the estimate, or more than 112 per cent of the estimate. This has a simple, definite meaning more significant for most practical purposes than a similar measure in terms of absolute values.[1]

[1] The significance of a measure of reliability in percentage form was pointed out by D. H. Davenport in 1922, in an unpublished article, and such a measure

To find the values of $2S$ or $3S$ these percentage figures may not be simply multiplied by 2 or 3. The value of $S_{\log y}$ must be so multiplied, and the resulting values reduced to natural numbers. For convenience in use, the anti-logarithms of both the positive and negative values should be secured, as in the preceding case. The computations are simple.

$$2S_{\log y} = .09802.$$

The anti-logarithm of this value, when considered positive is 1.25, when negative, .80.

$$3S_{\log y} = .14703.$$

The corresponding anti-logarithms are 1.40 and .71. Summarizing for the standard error, we have

$$S_r = .89 \text{ to } 1.12$$
$$2S_r = .80 \text{ to } 1.25$$
$$3S_r = .71 \text{ to } 1.40.$$

The values given for S_r indicate the probable percentage limits within which actual value and estimated value should fall in 68 out of 100 cases. The values given for $2S_r$ indicate the probable percentage limits in 95 out of 100 cases. The values of $3S_r$ indicate the probable percentage limits in 99.7 cases out of 100, always on the assumption of a normal distribution of the logarithms of the actual values about the fitted curve.

APPLICATION OF THE STANDARD ERROR OF ESTIMATE

We may illustrate the use of $S_{\log y}$. Given a production of oats 50 per cent above the trend value (i.e., the ratio to trend is 1.50), what is the most probable accompanying price ratio and what is the degree of accuracy of this estimate?

---

has been employed in several studies. There has not been available, however, a ready method of computing this measure, and its possibilities have not, therefore, been fully realized.

The estimating equation is

$$\log Y = (9.99139 - 10) - 1.18206 \log X.$$

Substituting in this equation the value .176091 (the logarithm of 1.50) we secure for log Y the value 9.78324 − 10. The corresponding natural number is .607. This means that if production is 150 per cent of normal (as measured by the given line of trend) price will probably be 60.7 per cent of normal (as measured by the line of trend).

To determine the reliability of this estimate, the standard error must be secured. Employing the values of S_r already computed we find that 54 is 89 per cent of 60.7, while 68 is 112 per cent of 60.7. We interpret these figures to mean that in 68 cases out of 100 the actual price prevailing under the given production conditions will not be less than 54 per cent of the normal or trend value nor more than 68 per cent of normal.[1] Corresponding values for $2S_r$ and $3S_r$ may be determined in the manner outlined above.

THE INDEX OF CORRELATION BASED ON LOGARITHMIC VALUES

We have still to compute the third measure, the abstract index of correlation.[2] For an equation of the type

$$\log Y = \log a + b \log X$$

the formula for ρ reduces to

$$\rho^2_{\log y \log x} = \frac{\log a \Sigma(\log Y) + b\Sigma(\log X \cdot \log Y) - Nc^2_{\log y}}{\Sigma(\log^2 Y) - Nc^2_{\log y}}$$

where $c_{\log y}$ represents the difference between the arithmetic mean of the logarithms of the Y-values and the origin (in this case, zero on the logarithmic scale). Substituting

[1] A question arises at once as to the adequacy of the given lines of trend, in the present problem. This question is discussed in greater detail in another section.

[2] The symbol ρ is used for this measure of correlation, instead of r, even though the relationship in logarithmic form is linear. This is done because such a measure, in terms of logarithms, cannot be interpreted in precisely the same way as the ordinary coefficient of correlation.

the proper values, we have

$$\rho^2_{\log y \log x} = \frac{(-.00861 \times -.32849) + (-1.18206 \times -.1143005) - (33 \times .00009909)}{.21721807 - (33 \times .00009909)}$$

$$= \frac{.13466882}{.2139481}$$

$$= .629445$$

$$\rho_{\log y \log x} = .793.$$

The index of correlation has a value of .793. How is this to be interpreted when we are dealing with logarithms as in the present case?

Its significance may be clearer if viewed in terms of the relationship

$$\rho^2_{\log y \log x} = 1 - \frac{S^2_{\log y}}{\sigma^2_{\log y}}.$$

In the present case these values are

$$S_{\log y} = .04901$$
$$\sigma_{\log y} = .08052.$$

When these values are squared and inserted in the above formula, we have

$$\rho^2_{\log y \log x} = 1 - \frac{.002402}{.006483}$$

and

$$\rho_{\log y \log x} = .793.$$

What does this value measure? We have seen that r and the more general index ρ are abstract measures of the degree of relationship between two variables, as this relationship is described by given functions. The value of ρ in a given case depends upon the variability about the fitted line, in relation to the variability about the mean of the Y's. If the variability of estimates is materially reduced when the equation of regression is used as a basis for estimates, instead of the mean Y, the equation may be assumed to describe a significant relationship. The value

of ρ depends thus upon the relation between the two quantities, S_y and σ_y.

In the cases dealt with in the preceding chapter the variability in each case was measured in terms of absolute deviations, and the value of ρ depended upon the relation between the two given measures of absolute variability. The sole difference in the present case is that we are working in terms of *logarithmic* or *ratio variability*, deviations being measured in terms of logarithms instead of natural numbers.

The index ρ must be interpreted in the light of this fact. Its value, as always, depends upon the relation between two measures of variability, S^2 and σ^2, but in the present instance these are expressed in terms of logarithms. In brief, the value of ρ depends upon the relation between the *ratio variability* about the fitted curve and the *ratio variability* about the geometric mean of the Y's. (It is the geometric mean of the Y's, because that is the value corresponding to the arithmetic mean of the Y logarithms.)

We have here a set of measures, therefore, which perform in the field of ratios precisely the same service as is performed in the field of natural numbers by S and ρ (in the linear case, r). These measures are secured in the same way as are S and ρ, except that the equation of relationship from which they are derived is one in which the dependent variable is $\log Y$ (or, in the reverse case, $\log X$). The general formulas for computing these values are the same as in dealing with natural numbers, except that $\log Y$ replaces Y throughout. The operation is analogous to that of using logarithmic paper instead of natural scale paper.

It should be noted that the values are in logarithmic or ratio form if Y is expressed logarithmically, whether X be so expressed or not. Thus we have fitted a curve of the type

$$\log Y = \log a + b \log X$$

the logarithmic form of the ordinary parabola or hyperbola.

The values S and ρ would also be in logarithmic form if the curve were of the type

$$\log Y = \log a + X \log b$$

the logarithmic form of the exponential

$$Y = a(b^X).$$

In each of these cases the logarithmic equation is linear, but this is not essential to the use of these measures. S and ρ are generally applicable measures, whether ratios or natural numbers be dealt with, and whether the functions be linear or otherwise.

It may be well at this point to summarize the symbols that have been used and to distinguish the different measures. We may employ the symbols S_y, σ_y, and ρ when arithmetic relations are in question, the two former being measures of variation in absolute terms, and the index ρ referring to degree of relationship when natural numbers are employed. If the logarithms of the Y's are used it is advisable to distinguish the symbols by subscripts, using $S_{\log y}$ and $\sigma_{\log y}$ as measures of the logarithmic variation about the fitted curve and about the arithmetic mean of the logarithms of the Y's, respectively. If $S_{\log y}$ is reduced to ratio form, it may be written S_r. Since the index ρ must be interpreted somewhat differently in this case, it may be written $\rho_{\log y \log x}$, or $\rho_{\log yx}$.

The Use of Reciprocals in the Measurement of Relationship

Another type of curve may be used to describe the relationship between the production and price of oats, and its use introduces us to a third field of correlation, a field in which somewhat new concepts enter, and in which the various measures must be interpreted in still another way.

This is a curve of the type

$$Y = \frac{1}{a + bX}$$

which may be expanded by adding additional terms to the denominator, as

$$Y = \frac{1}{a + bX + cX^2}.$$

This hyperbolic form has been used in several studies as an approximation to a "demand" curve for various commodities.

The equation to a curve of this type may be written

$$\frac{1}{Y} = a + bX$$

which is the equation to a straight line describing the relationship between the reciprocals of the Y's and the original X values. The normal equations required in fitting a curve of this type are

$$\text{I} \quad \Sigma\left(\frac{1}{Y}\right) = Na + b\Sigma(X)$$

$$\text{II} \quad \Sigma\left(\frac{X}{Y}\right) = a\Sigma(X) + b\Sigma(X)^2.$$

The method of computing the necessary values is illustrated in Table 134.

Substituting the proper values in the normal equations, we have

$$34.3360320 = 33a + 33.338b$$
$$35.2571485 = 33.338a + 34.168554b.$$

Solving,

$$a = -.1357$$
$$b = 1.1643.$$

TABLE 134

Computation of Values Required in Fitting a Curve to Data of Oat Production and Prices

EXAMPLE II

(1) Year	(2) Price Ratio Y	(3) Production Ratio X	(4) $\frac{1}{Y}$	(5) $\frac{X}{Y}$	(6) $\left(\frac{1}{Y}\right)^2$	(7) X^2
1881	1.30	.929	.7692308	.7146154	.59171602	.863041
1882	1.05	1.036	.9523810	.9866667	.90702957	1.073296
1883	.90	1.156	1.1111111	1.2844444	1.23456788	1.336336
1884	.85	1.128	1.1764706	1.3270588	1.38408307	1.272384
1885	.84	1.165	1.1904762	1.3869048	1.41723358	1.357225
1886	.77	1.108	1.2987013	1.4389610	1.68662507	1.227664
1887	.96	1.124	1.0416667	1.1708334	1.08506951	1.263376
1888	.79	1.151	1.2658228	1.4569620	1.60230736	1.324801
1889	.81	1.188	1.2345679	1.4666667	1.52415790	1.411344
1890	1.48	.798	.6756757	.5391892	.45653765	.636804
1891	1.10	1.088	.9090909	.9890909	.82644626	1.183744
1892	1.09	.943	.9174312	.8651376	.84168001	.889249
1893	1.16	.882	.8620690	.7603449	.74316296	.777924
1894	1.07	.886	.9345794	.8280373	.87343865	.784996
1895	.75	1.070	1.3333333	1.4266666	1.77777769	1.144900
1896	.76	.983	1.3157895	1.2934211	1.73130201	.966289
1897	.96	.969	1.0416667	1.0093750	1.08506951	.938961
1898	.95	1.005	1.0526316	1.0578948	1.10803329	1.010025
1899	.83	1.074	1.2048193	1.2939759	1.45158955	1.153476
1900	.86	1.033	1.1627907	1.2011628	1.35208221	1.067089
1901	1.37	.857	.7299270	.6255480	.53279343	.734449
1902	1.03	1.131	.9708738	1.0980583	.94259594	1.279161
1903	1.14	.911	.8771930	.7991228	.76946756	.829921
1904	.86	1.033	1.1627907	1.2011628	1.35208221	1.067089
1905	.86	1.090	1.1627907	1.2674419	1.35208221	1.188100
1906	1.04	1.013	.9615385	.9740385	.92455629	1.026169
1907	1.31	.770	.7633588	.5877863	.58271666	.592900
1908	1.29	.796	.7751938	.6170543	.60092543	.633616
1909	1.03	.978	.9708738	.9495146	.94259594	.956484
1910	.81	1.064	1.2345679	1.3135802	1.52415790	1.132096
1911	1.14	.810	.8771930	.7105263	.76946756	.656100
1912	.80	1.221	1.2500000	1.5262500	1.56250000	1.490841
1913	.87	.948	1.1494253	1.0896552	1.32117852	.898704
Total	32.83	33.338	34.3360320	35.2571485	36.85702940	34.168554

The desired equation is, therefore,

$$\frac{1}{Y} = -.1357 + 1.1643X.$$

THE STANDARD ERROR AND THE INDEX OF CORRELATION IN TERMS OF RECIPROCALS

To determine the utility of this equation we must have the standard error and the index of correlation. The two necessary formulas may be derived as in the preceding cases. Representing by $\frac{1}{Y}$ the reciprocal of an actual value we have, for each residual,

$$d = a + bX - \frac{1}{Y}. \qquad (1)$$

Multiplying by d and summing

$$\Sigma(d^2) = a\Sigma(d) + b\Sigma(dx) - \Sigma\left(\frac{d}{Y}\right).$$

Since

$$\Sigma(d) = 0 \text{ and } \Sigma(dX) = 0,$$

we have

$$\Sigma(d^2) = -\Sigma\left(\frac{d}{Y}\right). \qquad (2)$$

Multiplying the residual equation (1) now by $\frac{1}{Y}$, and summing, we have

$$\Sigma\left(\frac{d}{Y}\right) = a\Sigma\left(\frac{1}{Y}\right) + b\Sigma\left(\frac{X}{Y}\right) - \Sigma\left(\frac{1}{Y}\right)^2.$$

Substituting the equivalent of $\Sigma\left(\frac{d}{Y}\right)$ in the preceding equation (2), we secure

$$\Sigma(d^2) = \Sigma\left(\frac{1}{Y}\right)^2 - a\Sigma\left(\frac{1}{Y}\right) - b\Sigma\left(\frac{X}{Y}\right)$$

and for $S_{\frac{1}{y}}^2$, we have

$$S_{\frac{1}{y}}^2 = \frac{\Sigma\left(\frac{1}{Y}\right)^2 - a\Sigma\left(\frac{1}{Y}\right) - b\Sigma\left(\frac{X}{Y}\right)}{N}.$$

Inserting this value of $S_{\frac{1}{y}}^2$ in the general formula for the index of correlation

$$\rho^2 = 1 - \frac{S_{\frac{1}{y}}^2}{\sigma_{\frac{1}{y}}^2}$$

and simplifying, we have

$$\rho^2_{\frac{1}{y}x} = \frac{a\Sigma\left(\frac{1}{Y}\right) + b\Sigma\left(\frac{X}{Y}\right) - Nc_{\frac{1}{y}}^2}{\Sigma\left(\frac{1}{Y}\right)^2 - Nc_{\frac{1}{y}}^2}.$$

Inserting the proper values in these two equations, we find that

$$S_{\frac{1}{y}} = .1191$$
$$\rho_{\frac{1}{y}x} = .766.$$

For the standard deviation of the original Y-values, in terms of reciprocals, we secure

$$\sigma_{\frac{1}{y}} = .1851.$$

(The subscript $\frac{1}{y}$ is used in connection with each of these measures, as they should be distinguished from measures based upon natural numbers or logarithms.)

INTERPRETATION OF THE STANDARD ERROR OF ESTIMATE

How may we interpret these results? As in all former problems of this type the equation gives us a means of estimating Y from a known value of X. The standard error $S_{\frac{1}{y}}$ serves as a measure of the reliability of such estimates, and $\rho_{\frac{1}{y}x}$ is an abstract measure of the degree of relationship between the two variables. But in the present case all these measures are in terms of reciprocals. The equation enables us to estimate the reciprocal of Y, the

standard error has significance only in the form of a reciprocal, and the value of ρ depends upon the relation between two measures ($S_{\frac{1}{y}}^2$ and $\sigma_{\frac{1}{y}}^2$) both of which are in terms of reciprocals.

An illustration may make these meanings clear. If, in a given year, the production of oats is 150 per cent of trend, what is the most probable price? Substituting in the equation

$$\frac{1}{Y} = -.1357 + 1.1643X$$

a value of 1.50 for X, we have

$$\frac{1}{Y} = 1.6108$$

and

$$Y = .621.$$

We may expect a price approximately 62 per cent of trend. As a measure of the reliability of this estimate, we have

$$S_{\frac{1}{y}} = .1191.$$

This must be applied to the estimate in terms of reciprocals. Thus we have

$$1.6108 + .1191 = 1.7299$$
$$1.6108 - .1191 = 1.4917.$$

Reducing these reciprocals to natural numbers we secure .578 and .670 as the desired values. The most probable price, then, is 62.1 per cent of trend, and, on the assumption of an approximately normal distribution of reciprocals about the curve, the odds are 68 out of 100 that the price will fall between 57.8 per cent of trend and 67.0 per cent of trend. The limits of $2S$ and $3S$ may be similarly determined by adding to and subtracting from the estimate, as a reciprocal, amounts equal to twice .1191 and three times .1191. The results secured may then be converted

to natural numbers. Just as with logarithms, the value in absolute terms of a given difference between reciprocals varies at different points within the range of Y-values. Accordingly, the limits of reliability determined from $S_{\frac{1}{y}}$ should be expressed in natural numbers only after a particular estimate has been made.

A Comparison of Measures of Relationship

In interpreting ρ similar considerations enter. The value of the index of correlation, as we have seen, depends upon the degree of variation about the curve, as compared with the variation about the average of the original dependent series. In handling natural numbers, variability about the fitted line is compared with the variability about the *arithmetic mean* of the dependent variable, both measured in absolute terms (i.e., S_y is compared with σ_y). In handling logarithms, variability about the fitted line is compared with variability about the arithmetic mean of the logarithms of the dependent series, variability being measured in each case in terms of logarithms. But logarithmic deviations, as we have seen, may be interpreted in terms of ratios. The logarithmic deviations from the line represent the ratios of actual values to computed, while logarithmic deviations about the arithmetic mean of the logarithms of the original series represent the ratios of the actual values of the dependent series to their *geometric mean.* The value of $\rho_{\log y}$ depends upon the relation between these respective deviations (i.e., $S_{\log y}$ is compared with $\sigma_{\log y}$).

In fitting a curve in which the reciprocals of the dependent variable are employed, variability about the fitted line is measured in terms of reciprocals, and the variability of the original series is measured in the same terms. That is, $\sigma_{\frac{1}{y}}$ is computed from the differences between the reciprocals of the actual values and the arithmetic mean of all these reciprocals. But the arithmetic mean of these recipro-

cals is the reciprocal of the harmonic mean. Thus, in short, the value of the index of correlation, $\rho_{\frac{1}{y}}$, depends upon the relation between variability about the fitted line and variability about the *harmonic mean* of the dependent series, variation in both cases being measured in terms of reciprocals (i. e., $S_{\frac{1}{y}}$ is compared with $\sigma_{\frac{1}{y}}$).

We have, therefore, three broad families of curves for describing the relationship between variable quantities. These are:

1. Curves in the fitting of which natural values of the dependent variable are employed. Equations to all curves of this family will be of the type
$$Y = f(X).$$
2. Curves in the fitting of which logarithms of the dependent variable are employed. In all such cases the equations will be of the type
$$\log Y = f(X).$$
3. Curves in the fitting of which reciprocals of the dependent variable are employed. For these curves the equations will be of the type
$$\frac{1}{Y} = f(X).$$

In any one of these three cases the equations may be linear or non-linear. In so far as this problem of interpretation is concerned, there is no limitation as to the function of X which may be employed. (The computation of S and ρ by the methods suggested above involves certain limitations, which are outlined elsewhere.)

The standard error of estimate for the first family of curves is derived in terms of the original units of measurement (for the dependent variable) and has a direct and simple meaning in these terms. The index of correlation, for curves of this type, is a measure of the degree to which the *absolute variability* of the dependent variable may be

lessened by measuring deviations from the fitted curve instead of from the *arithmetic mean.*

The standard error of estimate for the second family of curves is derived, by the method outlined, in terms of logarithms. It is more convenient in general to give it meaning in terms of *ratios.* The index of correlation, $\rho_{\log y \log x}$, is a measure of the degree to which the logarithmic or ratio variability of the dependent variable may be lessened by computing deviations (or ratios) with the fitted curve instead of the *geometric mean* as base.

The standard error of estimate for the third family of curves is derived by the same process as in the other cases, but emerges as a *reciprocal.* The index of correlation, $\rho_{1_{\frac{1}{y}x}}$, is a measure of the degree to which the variability of the dependent variable, *in terms of reciprocals,* may be lessened by computing reciprocal deviations from the fitted curve instead of from the *harmonic mean.*

FACTORS GOVERNING THE CHOICE OF MEASURES OF RELATIONSHIP

It is clear, therefore, that the choice of a type of curve to describe a given relationship must be governed by basic considerations as to the type of average which is most appropriate as a measure of the central tendency of the given series. And this brings in a related question as to whether the dispersion about this average more nearly approximates the normal type when measured in absolute terms, in logarithms, or in reciprocals. In selecting a curve and in using the measures S and ρ there is always present an implicit assumption with respect to these points.

When absolute values are important, and the dispersion of the dependent variable approaches the normal type when plotted on an arithmetic scale, measures of relationship of the arithmetic type would appear to be appropriate. But, as we have seen, in handling series in which *rates* of change

rather than absolute amounts of change are of primary importance and the dispersion appears to follow a geometric law, the arithmetic mean and other arithmetic measures are notoriously inadequate. In such cases logarithmic curves seem preferable to arithmetic, and measures of the reliability of estimates and of degree of relationship which are based upon *ratios* seem to be more suitable than those based upon absolute values.

The harmonic mean has not been so widely employed as either of the above averages, and some attention may be given to principles governing its use in problems of the type here considered. In general, such harmonic measures are marked by the same weaknesses as the arithmetic, except that they err in the opposite direction. Geometric measures are perhaps better adapted to all-around employment than either. Yet in one particular field of interest to the economist the harmonic mean is particularly appropriate, and the utilization of reciprocals, as in the preceding example, seems to be justified.

The use of the harmonic mean assumes a normal distribution of reciprocals which, in natural numbers, means a much wider scatter above the average than below. The use of a curve of the type

$$\frac{1}{Y} = a + bX$$

involves a similar assumption as to the relation between Y and X. A given absolute increase in X will be accompanied by a certain decrease in the value of Y. The same absolute decrease in X will be accompanied by an increase in the value of Y which is larger than the decrease registered in the preceding case. But this is the relation which prevails, for many commodities, between the amounts produced and the price, the latter considered dependent. A given increase in production will cause some lowering of price. An equal decrease will cause a much greater increase in price. Moreover, when averaging the prices of such commodities

over a period, the harmonic mean may give a more typical value than any other average.[1] In such cases there is a strong *a priori* justification for using a curve of the reciprocal type and measuring the accuracy of all estimates in terms of harmonic relations.

ARITHMETIC, GEOMETRIC, AND HARMONIC MEASURES

The contrast between these different methods may be brought home most effectively by comparing the results obtained when curves of these three types are fitted to the same data. The computations involved in fitting curves of the second and third types (logarithmic and reciprocal) have been illustrated with reference to the data of oat production and prices (Table 132). A straight line (arith-

[1] "Buyers and sellers of potatoes are frequently mistaken as to the price justified by fundamental economic conditions. If such an error is general in the fall, it may happen, for example, that the price which results is too high. If the price is too high in the early part of the season, potatoes will not be consumed fast enough to dispose of the supply available. Farmers and dealers will then find that not all of the stocks on hand can be sold at existing prices. Since potatoes can not be carried over from one year to the next, the price, under such conditions as have been mentioned, must be lowered enough to permit the supply to be disposed of before the end of the season. A properly adjusted price would remain the same throughout the season, except for a gradual advance to cover cost of storage, and would maintain a fairly uniform consumption throughout the season. But since an abnormally high price early in the season causes small consumption, it must be compensated by an abnormally low price during the remainder of the season, or not all the crop can be sold.

"Similarly, if the price is abnormally low early in the season, the supply will be exhausted too rapidly and those who still have potatoes will find that they can get abnormally high prices for them during the remainder of the season."

But how, given the abnormally high or abnormally low prices during part of a season, may we compute the average price which would be justified by the true conditions of demand and supply, if these had been correctly estimated? Since "a low price during part of a season will be compensated only by a disproportionately high price during the remainder of the season" the arithmetic average for an entire season "will be somewhat higher than the average which would have resulted had a proper price been established at the beginning of the season. *This difficulty is eliminated by taking the harmonic mean of the monthly prices.*"

Holbrook Working, *Factors Determining the Price of Potatoes in St. Paul and Minneapolis.* Technical Bulletin 10, University of Minnesota Agricultural Experiment Station, 8–10.

metic) is fitted to the same data, and the necessary accompanying measures computed. The three sets of results are brought together in Table 135.

TABLE 135

Relation between the Production and Price of Oats, 1881–1913

Comparison of Results of Curve Fitting

(Prices are the dependent variable in each case)

	Equation	*Standard error of estimate*	*Index of correlation*
A	$Y = 2.24 - 1.236X$	$S_y = .12$	$r_{yx} = .783$
B	$\frac{1}{Y} = - .1357 + 1.1643X$	$S_{\frac{1}{y}} = .1191$	$\rho_{\frac{1}{y}x} = .766$
C	$\text{Log } Y = - .00861 - 1.18206 \log X$	$S_{\log y} = .04901$	$\rho_{\log y \log x} = .793$

It is impossible to compare the three standard errors as they stand, since only the first one is in the original units of measurement (ratio of actual price to trend). In the following table are given estimates, based on each of these equations, as to the most probable price (in terms of ratio to trend) which would accompany each of five different conditions of production.[1] Each estimate is accompanied by a series of values which indicate the limits set by the standard error. Throughout, the values of the estimates plus and minus S, $2S$, and $3S$ are given, in order to indicate the probable scatter of actual values about the estimates. The different amounts of variation which may be expected about each of the three lines of relationship are measured by the actual differences between the estimates and the limiting cases. These differences are given in the columns headed Δ. All values in this table are comparable, being reduced to the original units (ratio of actual price to trend).

[1] For the purpose of this illustration the limits of actual observation have been exceeded in setting up Table 136. Such extrapolation involves the possibility of errors of another sort. With these we are not here concerned.

TABLE 136

Comparison of Price Estimates and of Standard Errors of Estimate Based on Three Equations Relating to the Production and Price of Oats

(1) Value of X (ratio of production to normal)	(2) Estimated value of Y (ratio of price to trend) from arithmetic equation (A)	(3) Limits of arithmetic estimate	(4) Δ	(5) Estimated value of Y from reciprocal equation (B)	(6) Limits of estimate, reciprocal	(7) Δ	(8) Estimated value of Y from logarithmic equation (C)	(9) Limits of logarithmic estimate	(10) Δ
.5	1.622	+3S = 1.982	+.36	2.240	+3S = 11.223	+8.983	2.224	+3S = 3.114	+.890
		+2S = 1.862	+.24		+2S = 4.803	+2.563		+2S = 2.780	+.556
		+S = 1.742	+.12		+S = 3.055	+ 815		+S = 2.491	+.267
		−S = 1.502	−.12		−S = 1.768	−.472		−S = 1.979	− 245
		−2S = 1.382	−.24		−2S = 1.461	−.779		−2S = 1.779	−.445
		−3S = 1.262	−.36		−3S = 1.244	−.996		−3S = 1 579	−.645
.8	1 251	+3S = 1 611	+.36	1.257	+3S = 2.281	+1.024	1 276	+3S = 1.786	+ 510
		+2S = 1 491	+.24		+2S = 1.794	+.537		+2S = 1.595	+.319
		+S = 1.371	+.12		+S = 1.478	+.221		+S = 1 429	+.153
		−S = 1.131	−.12		−S = 1.093	−.164		−S = 1 136	−.140
		−2S = 1.011	−.24		−2S = .967	−.290		−2S = 1 021	−.255
		−3S = .891	−.36		−3S = .867	−.390		−3S = .906	−.370
1.0	1 004	+3S = 1 364	+.36	.972	+3S = 1.490	+.518	.980	+3S = 1.372	+.392
		+2S = 1.244	+ 24		+2S = 1.265	+.293		+2S = 1.225	+.245
		+S = 1.124	+.12		+S = 1.100	+.128		+S = 1.098	+.118
		−S = .884	−.12		−S = .871	−.101		−S = .872	−.108
		−2S = .764	−.24		−2S = .789	−.183		−2S = .784	−.196
		−3S = .644	− 36		−3S = .722	− 250		−3S = .696	−.284
1.2	.757	+3S = 1.117	+.36	.793	+3S = 1.106	+.313	.790	+3S = 1.106	+.316
		+2S = .997	+.24		+2S = .977	+.184		+2S = .987	+.197
		+S = .877	+.12		+S = .875	+.082		+S = .885	+.095
		−S = .637	−.12		−S = .724	− 069		−S = .703	−.087
		−2S = .517	−.24		−2S = .667	−.126		−2S = .632	−.158
		−3S = .397	− 36		−3S = .618	−.175		−3S = .561	−.229
1.5	.386	+3S = .746	+ 36	.621	+3S = .798	+.177	.607	+3S = 852	+.245
		+2S = .626	+.24		+2S = .728	+.107		+2S = .761	+ 154
		+S = .506	+.12		+S = .670	+.049		+S = .680	+.073
		−S = .266	−.12		−S = .578	− 043		−S = .542	−.065
		−2S = 146	−.24		−2S = 541	− 080		−2S = .484	−.123
		−3S = .026	− 36		−3S = 508	−.113		−3S = 433	−.174

ZONES OF ESTIMATE AND THEIR SIGNIFICANCE

A careful study of this table should make clear the nature of estimates based on the three types of equations here presented. The fundamental differences lie not so much in the actual values of the estimates, as in the standard errors which measure the reliability of these estimates and indicate the limits within which the actual values are likely

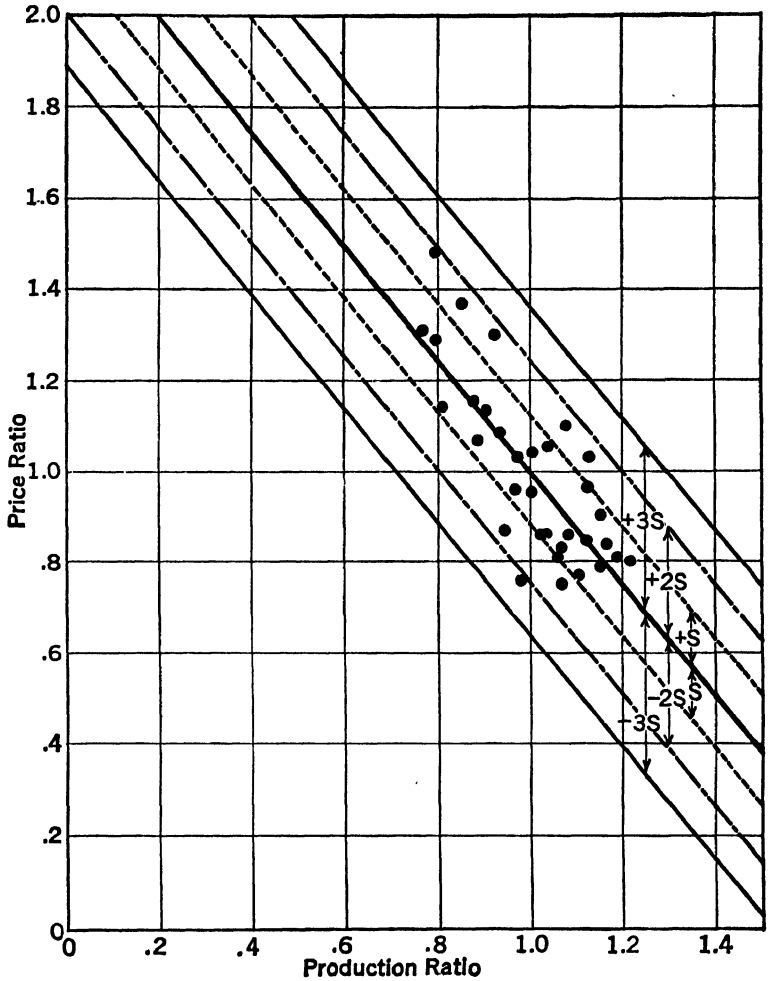


FIG. 87. — The Relation between the Production and Price of Oats: Illustrating the Use of an Arithmetic Equation of Regression and Arithmetic Zones of Estimate

to fall. In other words, the differences lie in the assumptions made as to the character of the scatter about the curves.

The measure S_y, which relates to the arithmetic curve, gives the same absolute range to errors of estimate whether the estimated value be high or low. An arithmetic dispersion

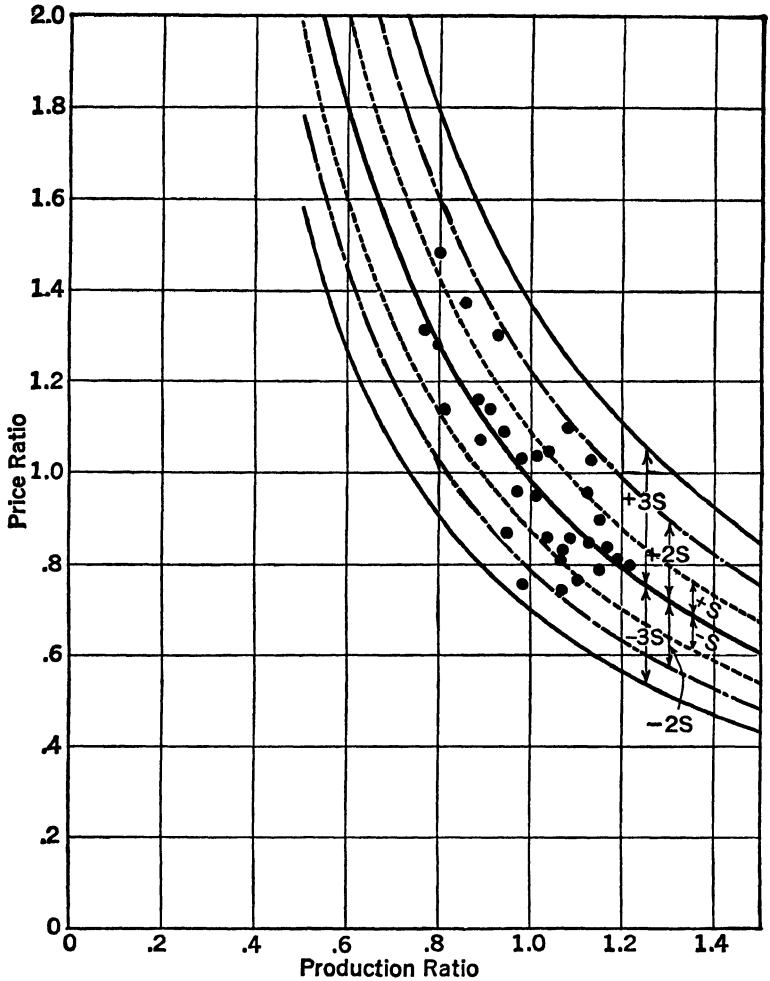


FIG. 88. — The Relation between the Production and Price of Oats: Illustrating the Use of a Logarithmic Equation of Regression and Geometric Zones of Estimate

about the curve is assumed. In each case the estimate is the arithmetic mean of the value which exceeds the estimate by an amount equal to S_y (or any multiple of S_y) and the value which falls below it by an equal amount. These conditions are brought out graphically in Fig. 87.

The original points are plotted, the straight line of relationship (arithmetic) is shown, and zones of estimate having

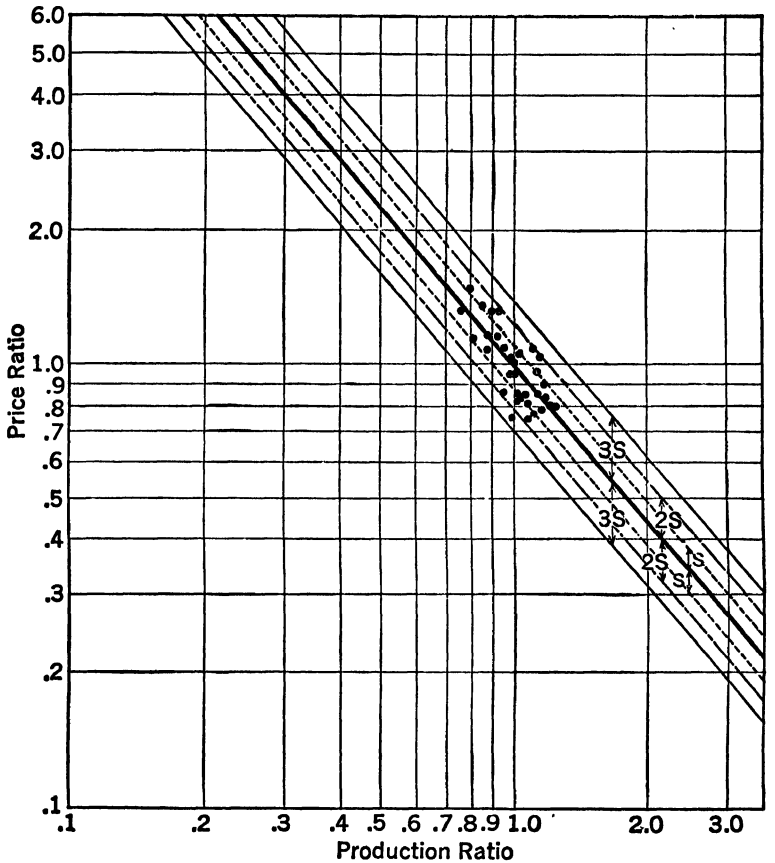


FIG. 89. — The Relation between the Production and Price of Oats: Illustrating the Use of a Logarithmic Equation of Regression and Geometric Zones of Estimate (Plotted on Double Logarithmic Paper)

widths, respectively, of $2S$, $4S$, and $6S$, centering at the fitted line, are marked out.

The measure $S_{\log y}$ gives the same relative or percentage range to errors of estimate, whether the estimate be high or low. This means that the absolute range within which the actual values should fall is much less when the estimates

are low than when they are high. It assumes a geometric dispersion about the curve which describes the relationship. The estimate is, in this case, the geometric mean of the value which exceeds it by an amount equal to $S_{\log y}$ (or any multiple of $S_{\log y}$) and the value which falls below it by an equal amount. Fig. 88 presents these relationships graphically. The original data are here plotted, together with the graph of the equation

$$Y = .9804X^{-1.18206}.$$

There are shown, also, the limits of zones of estimate having widths equal, respectively, to $2Sr$, $4Sr$, and $6Sr$, centering (geometrically) at the line of relationship. A comparison of Fig. 87 and Fig. 88 will reveal the differences between estimates based on the assumption of an arithmetic distribution and those based on the assumption of a geometric distribution.

The points and lines shown in Fig. 88 are plotted on a logarithmic scale in Fig. 89. On this scale the curve of relationship becomes straight, and the zones of estimate appear as symmetrical and of equal width throughout the range. This transformation when the data are plotted on logarithmic paper makes clear the fundamental simplicity of the assumptions involved in making estimates from logarithmic values.

In using the measure $S_{\frac{1}{y}}$ we carry still further the assumption that the variability about the curve is greater with high prices than with low. It shows a very limited range to errors of estimate when the estimate is low and a very wide range when the estimated price is high. A harmonic dispersion about the curve is assumed. The computed value, or estimate, is always the harmonic mean of the value which exceeds it by an amount equal to $S_{\frac{1}{y}}$ (or any multiple of $S_{\frac{1}{y}}$) and the value which falls below it by an equal amount.

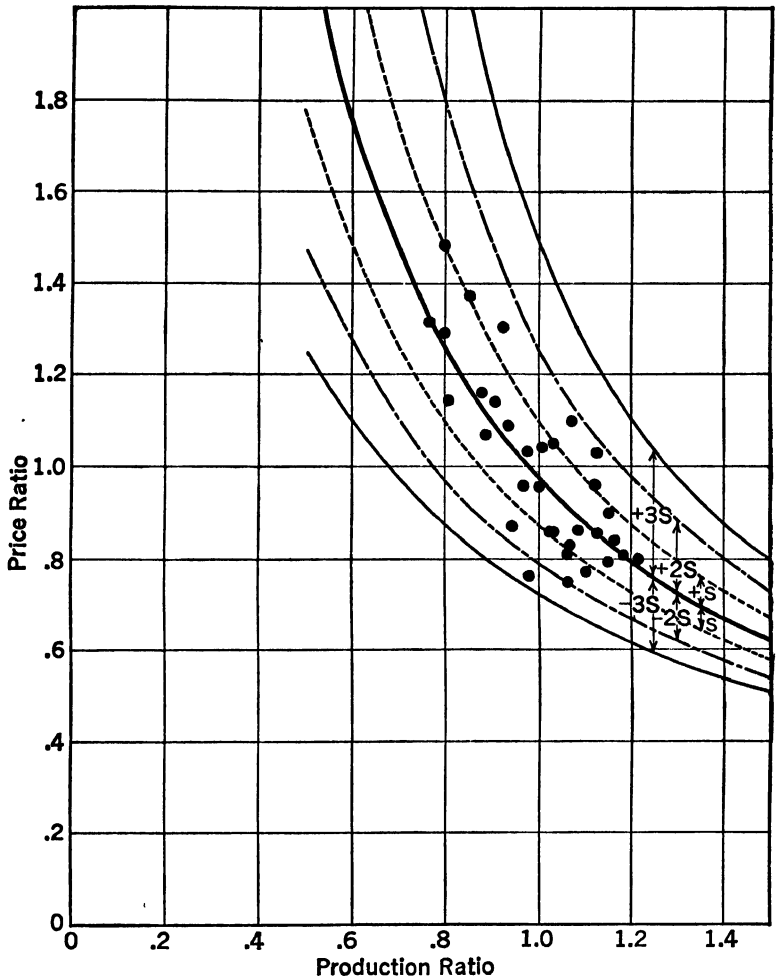


FIG. 90. — The Relation between the Production and Price of Oats: Illustrating the Use of an Equation of Regression Based upon Reciprocals, and of Harmonic Zones of Estimate

In Fig. 90 the curve $\frac{1}{Y} = -.1357 + 1.1643X$ is plotted, together with the original observations. Zones of estimate with widths of $2S_{\frac{1}{y}}$, $4S_{\frac{1}{y}}$, and $6S_{\frac{1}{y}}$, centering (harmonically) at the fitted line, are shown. The differences between this

figure and each of the two preceding are quite marked, particularly with respect to the zones of estimate. On the assumption of a normal harmonic distribution about the curve describing the relationship, the outer zone (with width equal to $6S$) marks the limits within which 99.7 per cent of all the points should fall, and the inner zone (with width equal to $2S$) marks the limits within which 68 per cent of all the points should fall. By plotting reciprocals throughout, instead of natural numbers, this apparently abnormal distribution could be reduced to the symmetrical form secured in plotting the geometric values on the logarithmic chart.

For both high and low estimates the geometric measure, $S_{\log y}$, stands between the arithmetic measure, S_y, and the harmonic measure, $S_{\frac{1}{y}}$. While the two latter have their particular functions, and are appropriate in certain cases, it is probably true that in using such methods as these in economic analysis, measures of the geometric family are more generally useful than those of the other types. This means, merely, that ratios are usually more important than absolute differences. It seems reasonable therefore to base estimates upon an equation of the type

$$\text{Log } Y = f(X)$$

and to measure the reliability of these estimates in terms of logarithms or ratios, using $S_{\log y}$ or S_r. In such cases, as we have seen, correlation is measured by $\rho_{\log y \log x}$ or $\rho_{\log yx}$. The value of this index depends upon the *ratio variability* about the curve, as compared with the *ratio variability* about the geometric mean.[1]

[1] The reasoning in C. M. Walsh's book, *The Problem of Estimation* (London, King, 1921, p. 12.) is peculiarly applicable to the present problem. Citing Galileo, in defence of the use of the geometric mean in averaging estimates, Walsh writes: "And so errors must be measured by an error *which is a ratio between the estimate and the true quantity, and not a concrete quantity itself.* We cannot measure errors by so many pounds, feet or crowns; we must measure them by the proportions of the pounds, feet or crowns in the erroneous estimates to the pounds, feet or crowns in the thing estimated." (Italics mine.) This ar-

REFERENCES

Killough, Hugh B., "What Makes the Price of Oats?" U. S. Department of Agriculture, *Bulletin No. 1351.*

Moore, H. L., "Elasticity of Demand and Flexibility of Prices," *Journal of the American Statistical Association*, March, 1922.

Moore, H. L., "Empirical Laws of Demand and Supply and the Flexibility of Prices," *Political Science Quarterly*, Dec., 1919.

Walsh, C. M., *The Problem of Estimation*, Chaps. 1–2.

Examples of curves employed to describe relationships of the type discussed in this chapter will be found in

Schultz, Henry, *Statistical Laws of Demand and Supply, with Special Application to Sugar.*

Warren, G. F. and Pearson, F. A., *Interrelationships of Supply and Price.*

Working, Holbrook, "Factors Determining the Price of Potatoes in St. Paul and Minneapolis," *Technical Bulletin 10*, University of Minnesota Agricultural Experiment Station.

---

gument bears out powerfully what has been said as to the use of logarithmic functions in estimating, and as to the employment of logarithmic measures of errors of estimate.

CHAPTER XVIII

STATISTICAL INDUCTION AND THE PROBLEM OF SAMPLING, CONCLUDED

The methods of induction discussed in an earlier section (Chapter XIV) dealt with the more familiar procedures employed in generalizing results secured from the study of samples. Certain research problems call for modifications of the methods there described, while for some purposes quite different instruments are needed. In the present chapter, therefore, we carry forward the discussion of statistical inference, considering methods appropriate to certain special conditions and special problems.

Generalizing from Small Samples

The standard error of an arithmetic mean, we have seen, is given by

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

where N is the number of observations in the sample and σ is the standard deviation of the population from which the sample is drawn. We do not know the standard deviation of the population but we approximate it from the standard deviation of the sample. (For convenience in this exposition we shall use s as a symbol for the standard deviation of the sample; σ will denote the standard deviation of the population.) This is an acceptable approximation when N is reasonably large, say 30 or more. But for small values of N the standard deviation of the sample is subject to a definite bias, tending to make it consistently lower than the standard deviation of the population. The value of σ_M derived by the customary method is also biased down-

ward. Therefore, when methods appropriate to large samples are employed with small samples, we consistently under-estimate the sampling errors to which our measurements are subject. This bias shows remarkable consistency, however. With samples of any stated size the magnitude of the error to be expected from the use of the standard deviation of the sample as an approximation to the standard deviation of the population may be determined, and correction made for it. Accordingly, generalization of results secured from small samples is possible. In the nature of things the margin of error in such generalization is larger than it is when large samples are used, but the distortion due to sheer bias may be avoided.[1]

The nature of the error involved in generalizing from small samples may be brought out in the following terms. If we represent by M the mean of the population from which a sample is drawn, by $\overline{X}$ the mean of a single sample, and by $\sigma_{\overline{x}}$ the standard deviation of a distribution of a number of $\overline{X}$'s computed from successive samples, we may write

$$T = \frac{\overline{X} - M}{\sigma_{\overline{x}}}.$$

The quantity T is the deviation of the mean of the sample from the mean of the population, expressed in units of the standard deviation of the sample means. When $\sigma_{\overline{x}}$ is determined from the actual distribution of a number of $\overline{X}$'s, or from the true standard deviation of the population and N of the sample, the quantity T may be interpreted as a normal deviate. The significance of given values of T may then be determined with reference to a table of areas under the normal curve. Actually, we do not have a large number of $\overline{X}$'s, which may be arranged in a frequency distribution, nor do we know the value of σ

[1] The bias involved in the use of s as an approximation to σ, for small samples, was first discovered by "Student." For the original memoir see "The Probable Error of the Mean," *Biometrika*, Vol. 6, 1908, 1–25.

(the standard deviation of the population), nor of $\sigma_{\bar{x}}$ (the standard error of $\bar{X}$). We approximate σ by s (the standard deviation of the sample) and $\sigma_{\bar{x}}$ by what we may call $s_{\bar{x}}$ ($s_{\bar{x}} = \frac{s}{\sqrt{N-1}}$ if s has been computed from $\sqrt{\frac{\Sigma d^2}{N}}$; $s_{\bar{x}} = \frac{s}{\sqrt{N}}$ if s has been computed from $\sqrt{\frac{\Sigma d^2}{N-1}}$). When these approximations are based upon small samples, the T derived from them may not be interpreted as a normal deviate. For the distribution of T varies with the size of the sample. With small samples the distribution departs significantly from the normal type. Statistical inferences that fail to take account of this are inaccurate.

A discussion in detail of the distributions of statistical measurements obtained from small samples would carry us beyond the scope of the present book. We may briefly note, however, certain characteristics of the distribution function of the standard deviation. These are effectively revealed by the results of an interesting experiment conducted by W. A. Shewhart.

Shewhart drew 1,000 samples, each consisting of four observations, from a normally distributed parent population with a known standard deviation, equal to unity.[1] The standard deviation, s, of each sample was computed. The distribution of these thousand values of s is represented by the dots in Fig. 91.[2] (The line running through the dots defines the theoretical distribution of s's to be expected, with samples of 4, on the basis of "Student's" theory. There is a notably close agreement between the theoretical and observed distributions.) Traditional sampling concepts would lead us to expect a normal distribution of s's, centering about 1, the value of σ in the parent population. Instead, the distribution is definitely skew, with the meas-

[1] W. A. Shewhart, *Economic Control of Quality of Manufactured Product*, New York, Van Nostrand, 1931, 163–173, 185–186.

[2] The figure is here reproduced with the permission of Dr. Shewhart and his publishers.

urements clustering about a central tendency well below unity. The mode of the thousand values of s here represented is, in fact, .717 and the arithmetic mean is .801. These s's, it will be recalled, represent estimates of σ.

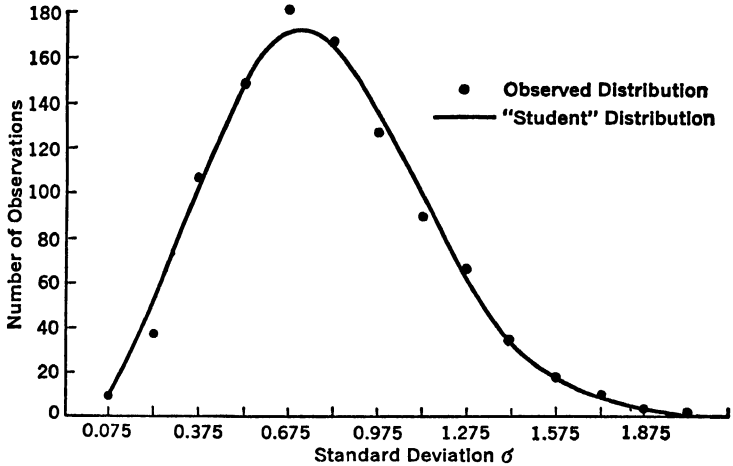


FIG. 91. — Distribution of Standard Deviations in Samples of Four Drawn from a Normal Universe

There is a clear tendency for such estimates, based on samples of four, to understate the true value.

The symbol T has been used above to define the deviation of a statistical measure from some standard or hypothetical value, expressed in units of the estimated standard error of the measure in question, when the deviation, so expressed, could be interpreted as a normal deviate. In the present exposition we shall employ the symbol t to relate to approximations to T when these approximations are based on small samples.

The difference between T and t may be reduced to more definite terms. If we let $x = \overline{X} - M$, we may write

$$T = \frac{x}{\sigma_{\bar{x}}}$$

$$t = \frac{x}{s_{\bar{x}}}.$$

We may derive t from T:

$$t = \frac{x}{\sigma_{\bar{x}}} \div \frac{s_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{x}{s_{\bar{x}}}.$$

The normally distributed quantity, $x/\sigma_{\bar{x}}$, has been divided by the factor $s_{\bar{x}}/\sigma_{\bar{x}}$, to give the quantity t. Opportunity to correct for the bias is given us, however, by the fact that the distribution of $s_{\bar{x}}/\sigma_{\bar{x}}$ is known. Thus the probability corresponding to any stated value of t may be determined (when t defines a departure from a certain hypothetical value, measured in units of $s_{\bar{x}}$).[1]

It is of some interest to compare values of t corresponding to stated probabilities, for samples of varying sizes, with values of T corresponding to the same probabilities. This is done in Table 137.

The familiar values given in the customary table of areas under the normal curve appear on the last line of

[1] The degree of error involved in using s as an approximation to σ, for small samples, is indicated by the following figures, taken from W. A. Shewhart (*loc. cit.*, 185). They define the relation between the modal s, for samples of size N drawn from a population of which the standard deviation is known, and the true σ of that population.

Size of sample N	*Modal s as a decimal fraction of true σ*
3	.577
4	.707
5	.775
6	.817
7	.845
8	.866
9	.882
10	.894
15	.931
20	.949
25	.959
30	.966
50	.980
100	.990

The fractions given above define relations that are to be expected on the basis of error theory, as modified by "Student" to take account of conditions affecting small samples. The modal value of the 1,000 standard deviations obtained by Shewhart in his empirical test of this theory was, as we have seen, .717 of the standard deviation of the universe. This result is very close indeed to the expected value of .707, for samples in which $N = 4$.

TABLE 137

Values of t and T Corresponding to Stated Probabilities [1]

n	Probability						
	.80	.50	.40	.20	.10	.05	.01
1	.325	1.000	1.376	3.078	6.314	12.706	63.657
2	.289	.816	1.061	1.886	2.920	4.303	9.925
3	.277	.765	.978	1.638	2.353	3.182	5.841
4	.271	.741	.941	1.533	2.132	2.776	4.604
5	.267	.727	.920	1.476	2.015	2.571	4.032
6	.265	.718	.906	1.440	1.943	2.447	3.707
7	.263	.711	.896	1.415	1.895	2.365	3.499
8	.262	.706	.889	1.397	1.860	2.306	3.355
9	.261	.703	.883	1.383	1.833	2.262	3.250
10	.260	.700	.879	1.372	1.812	2.228	3.169
20	.257	.687	.860	1.325	1.725	2.086	2.845
30	.256	.683	.854	1.310	1.697	2.042	2.750
∞	.25335	.67449	.84162	1.28155	1.64485	1.95996	2.57582

Table 137, for $n = \infty$. These are the values of T, as a normal deviate, corresponding to probabilities of .80, .50, etc. Thus, when we are dealing with infinitely large samples, the probability of a given sample yielding a value of T as great as .25335 or greater (either above or below the mean) is .80. (The area between the maximum ordinate and an ordinate erected at + .25335 is 10 per cent of the total area under the normal curve. Twenty per cent of the total area will fall within ± .25335, and 80 per cent will fall beyond these limits.) Similarly, just 50 per cent of the values of T will exceed the limits ± .67449; 5 per cent will exceed the limits ± 1.95996; 1 per cent will exceed the limits ± 2.57582.

As n grows smaller each of these limits must be extended, if the probabilities are to remain constant. For samples in which n is equal to 10, 50 per cent of the values of t will

[1] The entries in this table are extracts from a more detailed table (Table IV) in R. A. Fisher's *Statistical Methods for Research Workers*, Edinburgh, Oliver and Boyd, sixth edition, 1936. The table is printed here through the courtesy of Dr. Fisher and his publishers.

fall beyond the limits $\pm .700$; 5 per cent will exceed the limits ± 2.228, and 1 per cent will exceed the limits ± 3.169. (The letter n in Table 137 refers to the number of degrees of freedom in the computation of t. This general concept has been discussed in Chapter XV. When the arithmetic mean of a sample is being tested for significance, $n = N - 1$.) If in applying various statistical tests we attach significance to a given level of probabilities, such as 5/100 or 1/100, we must recognize that the values of t corresponding to these probabilities vary with n. Fortunately, we now know how these values vary and, using such a table as that given above, may make allowance for the variation.

For convenience in exposition we have distinguished T, as a normal deviate, from t, a similar deviate relating to a distribution of quantities derived from small samples, and therefore not normal. The probabilities corresponding to a given value of T are not the same as the probabilities corresponding to an identical value of t. Indeed, these probabilities vary for the same value of t computed from samples of different sizes. The distinction between T and t need not be preserved, however. We may use t generally to define the deviation of a statistical measure from some standard or hypothetical value, expressed in units of the standard error of the measure in question. The quantity t is to be interpreted as a normal deviate when large samples are dealt with. The interpretation is modified in dealing with small samples, as we have seen. The nature of the modification required is shown by the entries in Table 137 and in Appendix Table II.

EXAMPLES OF TESTS BASED ON t-TABLE

In determining whether the mean of a sample deviates significantly from any stated value we may compute t from the relation

$$t = \frac{\overline{X} - M}{s_{\bar{x}}}$$

where $\bar{X}$ is the mean of the sample, M is the stated value and $s_{\bar{x}}$ is an approximation to the standard error of $\bar{X}$. For this approximation we have

$$s_{\bar{x}} = \frac{s}{\sqrt{N}}$$

where s is the standard deviation of the sample (here computed from $\sqrt{\frac{\Sigma d^2}{N-1}}$). The value t, which for larger samples we have interpreted with reference to a table of areas under the normal curve, we here interpret with reference to the special t-table for small samples. In using the t-table for this purpose we take n of that table as equal to $N - 1$.

For the six New England states the average earnings of factory workers in 1935,[1] as indicated by census returns, were as follows:

Maine	$ 851
New Hampshire	892
Vermont	940
Massachusetts	1,007
Rhode Island	938
Connecticut	1,016
Average	$ 940.67

For s we obtain the figure $63.99. The standard error of the mean is $s_{\bar{x}} = \frac{s}{\sqrt{N}} = \frac{\$63.99}{\sqrt{6}} = \$26.13$.

Does the average of annual earnings of factory workers in the six New England states differ significantly from $1,022, the average for the country as a whole? Computing t we have

$$t = \frac{\bar{x} - M}{s_{\bar{x}}} = \frac{\$940.67 - \$1,022}{\$26.13}$$

$$= -3.11.$$

[1] These averages, and similar ones cited below, are derived by dividing the total wages paid by the average number of wage-earners employed during the year. Part-time workers are included. The averages do not represent full-time earnings, therefore.

Consulting the t-table with $n = 5$ we find that for $P = .01$, $t = 4.032$. The observed deviation is not as great as this. If our standard is a P of .01, the average for the New England states is not to be judged significantly less than the average for the country as a whole. If the standard were a P of .05, however, the deviation would be considered significant.

Similarly, we may test with reference to the t-table the significance of a difference between two means, computed from small samples. In this case we obtain t from the relation

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s}\sqrt{\frac{N_1 N_2}{N_1 + N_2}}$$

where the $\overline{X}$'s and N's have the customary meanings, and s is, in effect, an average standard deviation of the two distributions. For

$$s = \sqrt{\frac{\Sigma d_1^2 + \Sigma d_2^2}{N_1 + N_2 - 2}}.$$

Here d_1 and d_2 are used, respectively, to denote deviations of given observations from the means of the two distributions. The value t, as derived above, corresponds to $t = \frac{D}{\sigma_D}$, where D is the difference between two means and σ_D is the standard error of that difference. For small samples, however, the customary formula for σ_D is modified somewhat, and the special t-table rather than the table of normal deviates is used. In consulting the t-table in a problem of this type, n is taken as equal to $N_1 + N_2 - 2$.

Average earnings of workers employed in manufacturing plants in six Southern states, in 1935, are shown below:

North Carolina	$662
South Carolina	615
Georgia	599
Tennessee	744
Alabama	640
Mississippi	541
Average	$633.50

Does this average differ significantly from the mean earnings in six New England states in the same year? For the computation of s we have

$$s = \sqrt{\frac{20{,}491.3 + 23{,}153.5}{12 - 2}} = \$66.06$$

and for t

$$t = \frac{\$940.67 - \$633.5}{\$66.06}\sqrt{\frac{36}{12}}$$
$$= 8.05.$$

In the t-table, for $n = 10$, we find that the value of t corresponding to a P of .01 is 3.169. The present value is clearly significant. The two samples could not have come from one homogeneous parent population.

The t-table has particular value in connection with the interpretation of coefficients of regression. We may have observed that a given variable, Y, appears to increase by a constant increment or at a constant rate as another variable, x, changes in value. The degree of relationship between the two variables may be measured in terms of r, the coefficient of correlation, but special interest often attaches to the functional relationship and, in particular to the apparent regression of y on x. Does b of the equation of regression

$$Y = a + bX$$

depart significantly from zero, or from some other value which has significance for the purpose in mind? Here we must judge b with reference to the sampling errors to which it is exposed.

A general test of this type was applied in an earlier section (Chapter XIV), in seeking to determine whether average corn yield in Kansas had shown a significant decline over the period 1890–1933. For smaller samples we may compute t by exactly the methods there presented, but we should interpret t with reference to the special t-table

adapted to small samples. As a general formula we have

$$t = \frac{b - \beta}{\sigma_b}$$

where b is a coefficient of regression and β is a norm with reference to which we wish to judge the given value of b. For the standard error of b we have

$$\sigma_b = \frac{s_y}{\sqrt{\Sigma x^2}}$$

where

$$s_y = \sqrt{\frac{\Sigma(Y - Y_c)^2}{N - 2}}.$$

(In these expressions, $x = X - \overline{X}$, Y is an observed value of the dependent variable, and Y_c is the corresponding computed value.) In interpreting the value of t thus secured, the t-table is employed with $n = N - 2$.

This test may be extended to the comparison of two coefficients of regression. The series in Table 138 provide an illustration.

TABLE 138

Aggregate Values of Loans on Securities and Commercial Loans, Reporting Member Banks, Federal Reserve System, 1922–1929

(In hundreds of millions of dollars)

Year	*Loans on securities*	*Commercial loans ("all other loans")*
1922	39	73
1923	41	78
1924	45	80
1925	53	82
1926	57	86
1927	62	87
1928	69	89
1929	77	92

For loans on securities the trend (i.e., the equation of regression of volume of loans on time) is defined by

$$Y_1 = 30.63 + 5.49X_1.$$

The corresponding equation for commercial loans is

$$Y_2 = 72.13 + 2.54X_2.$$

In each case the origin is at 1921. The eight-year period was marked by an increase of loans on securities which was much more rapid than the corresponding advance in commercial loans. We must ask, however, whether the difference between the two coefficients of regression is really significant, if account be taken of sampling fluctuations.

The coefficients to be compared are

$$b_1 = 5.49$$
$$b_2 = 2.54.$$

In testing whether $b_1 - b_2$ is significant (i.e., deviates significantly from zero) we must compute

$$t = \frac{b_1 - b_2}{\sigma_{b_1 - b_2}}$$

$\sigma_{b_1-b_2}$ being, of course, the standard error of the difference between the two coefficients of regression. For this standard error we have

$$\sigma_{b_1-b_2} = \sqrt{\frac{S_y^2}{\Sigma(x_1^2)} + \frac{S_y^2}{\Sigma(x_2^2)}}$$

where x_1 and x_2 are given values of the two variables, expressed as deviations from their respective arithmetic means, and

$$S_y^2 = \frac{\Sigma(Y_1 - Y_{c1})^2 + \Sigma(Y_2 - Y_{c2})^2}{N_1 + N_2 - 4}.$$

S_y^2 is a measure of the average scatter about the two lines of regression.

In the present example we have

$$S_y^2 = 2.40$$

$$\sigma_{b_1-b_2} = \sqrt{\frac{4.80}{42}} = .338$$

$$t = \frac{b_1 - b_2}{\sigma_{b_1-b_2}} = \frac{5.49 - 2.54}{.338} = 8.73.$$

For the interpretation of this value of t we enter the t-table with $n = N_1 + N_2 - 4 = 12$. In this case the value of t far exceeds the value of 3.055, corresponding to $P = .01$. The results are not consistent with the hypothesis that the true value of $b_1 - b_2$ is zero. The trends of the two series differ significantly. (Here, again, the reader should bear in mind that such tests of significance apply only with important qualifications to economic series that are ordered in time.)

Sampling Errors of Coefficients of Correlation Computed from Small Samples

As a general formula for the determination of the standard error of the coefficient of correlation we have made use of

$$\sigma_r = \frac{1 - r^2}{\sqrt{N - 1}}.$$

In error theory, the r that appears in the numerator of the right-hand member of this equation is the coefficient of correlation in the universe from which the sample in question is drawn. But this r is not known. Our best approximation to it is the r derived from the sample. Here, again, we face distortion in small samples, a distortion that is the greater the higher the value of the true correlation. The nature of this bias may be readily understood. If we are drawing samples from a universe in which the true value of r is $+ .95$, the range of the possible variation of the sample r's above the true r is only .05. But the range of possible variation below the true value is 1.95 (i.e., from $+ .95$ to $- 1.00$). Accordingly, a distribution of r's obtained from a great many small samples from this universe will be sharply skew. An estimate of the true value based upon a sample value will be subject to corresponding bias. This bias will not be present when the population value of r is zero. (The distribution of sample r's when the population value of r is zero will be symmetrical,

but will depart somewhat from the normal type in other respects.) It will not be pronounced when the samples are large, even for high values of r. But when samples are small and the population value of r departs materially from zero, substantial inaccuracy results from the use of the formula given above.

Allowance may be made for this bias by use of the table showing the distribution of t, for samples of various sizes. R. A. Fisher has shown that the procedure employed in deriving t, in testing whether a coefficient of linear regression differs significantly from zero, may be used, with an algebraic modification of the mathematical expression, in determining the significance of r. If we are testing the hypothesis that a sample from which a given r has been computed was drawn from a population in which the true value of r is zero, we may compute t from the relation

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}.$$

This is equivalent, of course, to dividing the quantity $r - 0$ (i.e., the deviation of the given r from the hypothetical value of zero) by $\sqrt{1-r^2}/\sqrt{N-2}$. In consulting the t-table for the interpretation of the values thus obtained, n, the number of degrees of freedom, is taken as equal to $N - 2$.

As an illustration, we may test the results obtained from a study of the relation between the production and the price of cotton in the United States, covering 35 observations. The value of r is $-.65$. We have

$$t = \frac{-.65\sqrt{35-2}}{\sqrt{1-(-.65)^2}} = -4.91.$$

In consulting the t-table we find that for $n = 33$ the value of t corresponding to a probability of 1 per cent [1] is approxi-

[1] This probability refers to the likelihood of deviations above or below the assumed true value of zero. It corresponds to the sum of areas at both extremities of a frequency curve. We may divide it by two to obtain the probability of a deviation of the stated magnitude in one direction only from the hypothet-

TABLE 139

Values of the Correlation Coefficient for Different Levels of Significance[1]

n	$P = .05$	$P = .02$	$P = .01$
1	.996917	.9995066	.9998766
2	.95000	.98000	.990000
3	.8783	.93433	.95873
4	.8114	.8822	.91720
5	.7545	.8329	.8745
6	.7067	.7887	.8343
7	.6664	.7498	.7977
8	.6319	.7155	.7646
9	.6021	.6851	.7348
10	.5760	.6581	.7079
11	.5529	.6339	.6835
12	.5324	.6120	.6614
13	.5139	.5923	.6411
14	.4973	.5742	.6226
15	.4821	.5577	.6055
16	.4683	.5425	.5897
17	.4555	.5285	.5751
18	.4438	.5155	.5614
19	.4329	.5034	.5487
20	.4227	.4921	.5368
25	.3809	.4451	.4869
30	.3494	.4093	.4487
35	.3246	.3810	.4182
40	.3044	.3578	.3932
45	.2875	.3384	.3721
50	.2732	.3218	.3541
60	.2500	.2948	.3248
70	.2319	.2737	.3017
80	.2172	.2565	.2830
90	.2050	.2422	.2673
100	.1946	.2301	.2540

ical value. In most problems of the type here discussed it is conservative practice to test given results with reference to the probability of a deviation of given magnitude, without consideration of the direction of deviation. The tabulated values of t lend themselves to this procedure.

[1] This table is printed here through the courtesy of R. A. Fisher and his publishers, Oliver and Boyd, of Edinburgh. The original appears as Table V.A of *Statistical Methods for Research Workers*.

mately 2.73. If the true value of t were zero, a value as great as 2.73 or greater would occur only 1 time out of 100, as a result of chance fluctuations of sampling. The present value of t is substantially greater than 2.73. It it highly improbable that it reflects a chance drawing from a population in which the true value of t (and, of course, of r) is zero. The results we have obtained are not, then, consistent with the hypothesis that the true value of r is zero. There appears to be a significant negative correlation between the production and the price of cotton.

If we are seeking to determine the significance of given coefficients of correlation with reference to hypothetical values of zero, use may be made of a table prepared by R. A. Fisher, showing the values of correlation coefficients at stated levels of significance. Selected values from this table are given in Table 139 and in Appendix Table III. In simple correlation problems, this is to be read with n equal to $N - 2$ (the number of pairs of original observations less 2). In determining the significance of coefficients of partial correlation the number of variables held constant is also subtracted from N.

The use of the table requires little explanation. If a sample is based on 12 pairs of observations, with n equal to 10, we would require a coefficient at least as high as .7079 before we accept it as significant, if our standard of significance is $P = .01$. For only 1 time out of 100 trials would a sample of 12 drawn from an uncorrelated population yield a value of r as great as .7079. If our standard of significance is $P = .05$ we would accept as significant of a real relationship an r of .5760, or greater, obtained from a sample of 12.

TRANSFORMATION OF r TO z

The sampling limitations attaching to r have led R. A. Fisher to utilize as a general measure of linear correlation a loga-

rithmic function of r that possesses certain distinctive merits.[1] In effecting the transformation we have

$$z = \tfrac{1}{2}\{\log_e(1 + r) - \log_e(1 - r)\}.$$

Conversely

$$r = (e^{2z} - 1) \div (e^{2z} + 1).$$

The scales of possible values of r and z are, of course, quite different. For $r = 0$, $z = 0$, and for $r = 1$, $z = \infty$. Negative values of r give negative values of z. The relations between the two functions, at different levels of correlation, are shown by the entries in Appendix Table IV. Transformation may be more readily effected by means of this table than from the relations given above.

There are certain highly important advantages in this transformation. Not least is the replacing of r by a function with a distribution of values corresponding more closely to the true significance of observed correlations than do those of r. Thus a change in the value of r from .88 to .98 is equivalent, on the r scale, to a change from .20 to .30. But the first of these differences represents, on the z scale, a change from 1.38 to 2.30 (a range of .92) while the second represents a change in z from .20 to .31 (a range of .11). The first difference, on the z scale, is over 8 times more significant than the second. In this the z scale gives a far more accurate representation of the true significance of observed correlations than does the r scale.

More important than this, however, is the fact that the distribution of z is much closer to the normal type than is that of r; in particular, the distribution of z is not subject, as is that of r, to marked variations in form with variations in the degree of correlation in the population. The form of the distribution of z is virtually independent of the degree of correlation. As a result, the sampling errors to which z is exposed may be estimated with considerable

[1] See *Statistical Methods for Research Workers*, Chapter VI.

accuracy. For the standard error of z we have

$$\sigma_z = \frac{1}{\sqrt{N-3}}.$$

This standard error, it is to be noted, is a function solely of N. It is independent of the true value of z in the parent population.

From the example in Chapter XVI we obtained a coefficient of partial correlation of $-.2923$ between corn yield per acre in Kansas and average June temperature, holding constant effects of changes in July and August temperatures. Referring to Appendix Table IV we have, for $r = -.2923$, $z = -.301$. In computing the standard error of a coefficient of partial correlation we must subtract from N the number of variables held constant. Since N equals 44 in the example in question, we treat the coefficient of partial correlation as we would a simple coefficient based on 42 observations. For the standard error of z we have, then,

$$\sigma_z = \frac{1}{\sqrt{42-3}} = .160.$$

With reference to this result we may determine whether z differs significantly from zero. For the test we must have

$$\frac{z-0}{\sigma_z} = \frac{-.301}{.160} = -1.88.$$

We interpret 1.88 as a normal deviate. It is clear that it is not large enough to indicate that z is significant. The result is not inconsistent with the hypothesis that the true value of z (and hence of r) is zero.

If, however, we test the coefficient $r_{14.23} = -.4057$, from the same example (defining the relation between corn yield per acre and August temperature, with June and July temperatures held constant), we have

$$\frac{z-0}{\sigma_z} = \frac{-.430}{.160} = -2.69.$$

This result is clearly significant. So, also, is the measure $r_{13.24} = -.6101$, the coefficient of partial correlation between corn yield and July temperature, with June and August temperatures held constant.

The procedure would be similar, of course, if we were testing the significance of the deviation of an observed value of z from a theoretical value other than zero.

The transformation to z makes possible, also, an accurate test of the significance of the difference between two observed correlations. The standard error of the difference between two values of z is given by

$$\sigma_{D_z} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$$

where N_1 is the number of pairs of observations in the first sample, N_2 the number in the second.

This test may be illustrated with reference to observations on the timing of price changes during business cycles. For 111 commodities we have observations on the timing of price declines in two successive periods of business recession occurring in the late 90's and early 1900's. The degree of relation between the time sequences of commodity price changes in these two recessions is indicated by a coefficient of correlation of $+.22$. For two similar (successive) periods in the 1920's the measure of correlation, based on the prices of 121 commodities, has a value of $+.36$. There appears to have been a closer approach to a common pattern in the later period than in the earlier. In testing the significance of the difference between the two results we set up the hypothesis that the two samples were drawn from the same parent population, and that therefore the true value of the difference between the two coefficients is zero.

For the two samples we have

$$r = .22;\ z = .223;\ \frac{1}{N_1 - 3} = \frac{1}{108} = .0093$$

$$r = .36;\ z = .377;\ \frac{1}{N_2 - 3} = \frac{1}{118} = .0085.$$

The difference to be tested is

$$D_z = .377 - 223 = .154.$$

The standard error of this difference is

$$\sigma_{D_z} = \sqrt{.0093 + .0085} = .133.$$

We wish to know whether Dz is significantly different from zero. We compute, therefore,

$$\frac{D_z - 0}{\sigma_{D_z}} = \frac{.154 - 0}{.133} = 1.16.$$

Interpreting 1.16 as a normal deviate, we conclude that the difference is not significant. Dz differs from the hypothetical value of zero by only slightly more than one standard deviation. The results are not inconsistent with the hypothesis that the two samples are drawings from the same parent population. There is here no clear evidence that the degree of relationship between price movements in successive cycles was closer in the 1920's than in the earlier period.[1]

Finally, making use of the z-transformation, we may combine results secured from the measurement of correlation in different samples. If we have two values of r, obtained from samples drawn from the same population, a weighted average of the two will provide a better estimate of the true correlation than will either of the r's, taken separately. For the averaging process we transform the r's to z's, weight each z by the corresponding N, less 3, and average them. Then, if desirable, the corresponding value of r may be determined. We may note that the

[1] The time factor enters to cloud statistical inductions relating to samples drawn from different periods (see above, Chapter XIV). Such an induction should be supported by evidence indicating that fundamental conditions in the field in question have not been altered over the time interval involved. This caution does not, of course, affect the procedure illustrated above.

standard error of the weighted average of the two z's is given by

$$\sqrt{\frac{1}{(N_1 - 3) + (N_2 - 3)}}.$$

The Chi-Square Test

One of the great contributions of Karl Pearson to statistical methodology was the determination of the form of the distribution of Chi-square, and the development of methods of utilizing this distribution. The character of this distribution and various tests based on it are our concern in the present section.

THE NATURE OF CHI-SQUARE AND ITS DISTRIBUTION

The quantity Chi-square (represented always by the symbol χ^2) is a measure of the degree to which a series of observed frequencies deviate from corresponding theoretical or hypothetical frequencies. The theoretical frequencies are set up on the basis of some hypothesis, some rational argument. The magnitude of the discrepancy between theory and observation is defined by the quantity χ^2. It was Pearson's contribution to determine the nature of the distribution of the values of χ^2 that would be obtained under given sampling conditions. Knowledge of this distribution enables us to determine whether a given discrepancy between theory and observation may be attributed to chance, or whether it results from the inadequacy of the theory to fit the observed facts. This instrument is obviously one of extreme importance in statistical analysis.

The character of the distribution of χ^2 may be discussed with reference to Weldon's data relating to the results obtained in 4,096 throws of 12 dice (see page 433). We call a 4, 5, or 6 spot a success, a 1, 2, or 3 spot a failure. When 12 dice are thrown the expected (or theoretical) number of successes on each throw is 6. A deviation from

6 represents a discrepancy between expectation and observation. From the result of each throw of 12 dice a value of χ^2 may be computed. Thus, a given throw yields 2 successes and 10 failures. The 2 successes represent a deviation of 4 from the expected value of 6; the 10 failures represent a deviation of 4 from the expected value of 6. (In such an experiment as this there are two components of each value of χ^2, even though when one component is given the other is necessarily determined. For the sum of successes and failures must be 12 on each throw.) The value of χ^2 in a given instance is obtained by squaring the discrepancies between expectation and observation, dividing the squared values by the corresponding expected values, and adding the quantities thus obtained. That is

$$\chi^2 = \Sigma\left\{\frac{(f_0 - f)^2}{f}\right\}$$

where f_0 denotes an observed frequency and f defines the corresponding theoretical frequency.

In the case cited above we have

$$\chi^2 = \frac{(2-6)^2}{6} + \frac{(10-6)^2}{6} = 5.333.$$

On another trial, with 7 successes and 5 failures, we have

$$\chi^2 = \frac{(7-6)^2}{6} + \frac{(5-6)^2}{6} = .333.$$

On still another trial, giving 6 successes and 6 failures, we have

$$\chi^2 = \frac{(6-6)^2}{6} + \frac{(6-6)^2}{6} = 0.$$

The 4,096 throws thus yield 4,096 values of χ^2. Tabulating these with respect to the frequency of occurrence of stated values, we obtain the distribution given in Table 140 on page 620.

This table gives us information as to the nature of the discrepancies between theoretical norms and actual results

TABLE 140

Tabulation of 4,096 *Observed Values of* χ^2

Value[1] *of* χ^2 *(measuring deviation of observation from expectancy in dice-throwing experiment)*	*Frequency of occurrence (absolute)*	*Frequency of occurrence (relative)*
0 to .833	2,526	.6167
.833 to 2.167	966	.2358
2.167 to 4.167	455	.1111
4.167 to 6.667	131	.0320
Over 6.667	18	.0044
Total	4,096	1.0000

that chance may bring about. For deviations from the expected frequency of successes, 6, may be attributed to the mass of undifferentiated causes we call chance. The magnitude of χ^2 varies, of course, with the degree of deviation. Values of χ^2 not exceeding .833 are most frequent. Higher values of χ^2 occur with decreasing frequency. Only 18 out of 4,096 observed values of χ^2 exceed 6.667. This distribution furnishes us, therefore, with a standard of reference to employ when seeking to determine whether a given discrepancy between theoretical and observed values is attributable to chance, or whether it is too great to be so explained.

This use of the table, as a standard for determining the probability that given discrepancies between theory and observation are attributable to the play of chance, is facilitated by a somewhat different arrangement. We may set up a table of cumulative values, based upon the

[1] The 4,096 values of χ^2 tabulated here constitute a discrete series. The conditions of the experiment are such that the 4,096 observations on χ^2 are distributed among only seven values, ranging from 0 to 12. In order that the observed frequencies of occurrence of stated values of χ^2 may be compared (in a later table) with theoretical frequencies, an uneven class-interval is employed above. Class limits are taken midway between successive values at which the actual observations fall. (The decimal fractions used in the table do not define these limits with full accuracy.)

tabulation of the 4,096 values of χ^2 obtained in the preceding experiment. These are given in Table 141.

The entries in col. (2) of this table indicate that in the experiment involving 4,096 throws of dice, a value of χ^2 of 6.667 or more occurs less frequently than 1 time out of 100 (only 44 times out of 10,000, in fact). A value as great as 4.167, however, occurred more frequently than 3 times out of 100. If we interpret these relative frequencies as probabilities, we may obtain from such a table a knowledge of the probabilities corresponding to stated values of χ^2. Here is the instrument we desire, in seeking to determine whether given observations conform closely enough with expectations based on theory, or on working hypotheses which perhaps are not yet ready to be dignified as theories.

TABLE 141

Cumulative Relative Frequencies of Occurrence of 4,096 *Observed Values of* χ^2, *with Corresponding Theoretical Frequencies* [1]

(1) *Value of* χ^2 *(cumulative deviation of observation from expectancy)*	(2) *Relative frequency of occurrence (observed)*	(3) *Relative frequency of occurrence (theoretical)*
0 or more	1.0000	1.0000
.833 or more	.3833	.3613
2.167 or more	.1475	.1411
4.167 or more	.0364	.0412
6.667 or more	.0044	.0098

We should note two important limitations attaching to the entries in col. (2) of the above table, showing relative frequencies corresponding to stated values of χ^2. In the first place, these are merely empirical results, obtained from a given set of experiments. The conditions of the experiment yield a discontinuous series of values for χ^2. In some degree, this discontinuity has been ironed out by

[1] One degree of freedom is present in the determination of a single value of χ^2, in this example

the method of classification employed, but the instrument derived from this single experiment remains an imperfect one. The effects of chance fluctuations are present in these results, also, and contribute to the imperfection of the instrument. The true distribution of χ^2 is only approximated by the results presented in col. (2) of Table 141.

The entries in col. (3) of Table 141 are free of this limitation. These record the frequencies with which values of χ^2 falling within the limits indicated in col. (1) might be expected to occur, on the basis of mathematical theory, under the conditions of the present experiment.[1] These are the entries which provide the standard we desire, in determining the significance of a given series of discrepancies between observation and expectation. It is to be noted, however, that the empirically derived table constitutes a fair approximation to the theoretical distribution of χ^2 under these conditions.

The second limitation attaching to the example cited above is that each of the 4,096 values of χ^2 tabulated has two components, and that the experiment is such that when one component is given the second is necessarily determined. (Since there are 12 events in each throw we know, for example, that if we have 8 successes there must be 4 failures.) This condition is described by saying that there is but one degree of freedom in the derivation of a given value of χ^2. The table we have obtained relates, therefore, to a special case — the distribution of values of χ^2 computed with one degree of freedom. There are other possible cases. For each of these the distribution of χ^2 may be determined in a manner similar to that shown above.

As an example of a different set of conditions we may consider the outcome of a throw of 24 dice, account being kept of the frequency of occurrence of each possible result

[1] These relative frequencies are taken from G. Udney Yule "Table of the values of P for divergence from independence in the fourfold table," *Journal of the Royal Statistical Society*, Vol. LXXXV, January, 1922, 103–104.

(i.e., the appearance of a 1, 2, 3, 4, 5, or 6 spot). When 24 dice are thrown there may be expected 4 one spots, 4 two spots, 4 three spots, etc. In a given throw we obtain the following results:

	Number of spots					
	1	2	3	4	5	6
Observed frequency	2	5	6	4	4	3
Expected frequency	4	4	4	4	4	4

For the results of this throw the value of Chi-square would be given by

$$\chi^2 = \frac{(2-4)^2}{4} + \frac{(5-4)^2}{4} + \frac{(6-4)^2}{4} + \frac{(4-4)^2}{4} + \frac{(4-4)^2}{4} + \frac{(3-4)^2}{4} = 2.50.$$

This quantity has six components. However, as soon as five are given the sixth is determined, since the total number of events is fixed at twenty-four. There are, then, five degrees of freedom in the calculation of χ^2 in this experiment.

If the 24 dice were thrown a thousand times, say, we should have one thousand values of χ^2. A distribution of these could be constructed, similar to that derived empirically for the case in which there was one degree of freedom. It would be a different distribution, however, for the change in degrees of freedom has an obvious relation to the magnitude of χ^2. The character of the distribution of the values of χ^2 that would be obtained in such an experiment is indicated by the entries in Table 142 on page 624. We do not here give empirical values, as in the preceding example. The table shows the theoretical frequencies with which given values of χ^2 occur, when five degrees of freedom prevail.

In using tables of this sort we may interpret measures of relative frequency as probabilities. Thus we may read Table 142, which relates to the distribution of χ^2 computed with five degrees of freedom, as follows: If

TABLE 142

Tabulation of χ^2 Computed with Five Degrees of Freedom, with Cumulative Relative Frequencies [1]

Value of χ^2	*Relative frequency of occurrence (theoretical)*
0 or more	1.0000
1 or more	.9626
2 or more	.8491
3 or more	.7000
4 or more	.5494
5 or more	.4159
6 or more	.3062
7 or more	.2206
8 or more	.1562
9 or more	.1091
10 or more	.0752
11 or more	.0514
12 or more	.0348
13 or more	.0234
14 or more	.0156
15 or more	.0104
16 or more	.0068
30 or more	.000015
∞	.000000

the true value of χ^2 is zero (i.e., in an infinitely large sample observed frequencies would agree precisely with the theoretical frequencies we have set up), the probability of our securing a χ^2 of zero or more, from a sample of the type here employed, is 1.00; the probability of our securing a χ^2 of 1.00 or more is 9,626/10,000; the probability of our securing a χ^2 of 3.00 or more is 7/10; the probability of our securing a χ^2 infinitely large is 0. The quantities χ^2 and P stand, thus, in a definite functional relationship, for any given value of n (n denotes the number of degrees of freedom). At the two limits the relationships are the

[1] From the table prepared by W. P. Elderton and given in *Tables for Statisticians and Biometricians*, Karl Pearson, editor, 26. The n' of Elderton's table is equal to $n + 1$, for an example of the type here given.

TABLE 143 [1]

Table of χ^2 for Selected Values of P and n

n	$P = .99$	.95	.50	.10	.05	.02	.01
1	.000157	.00393	.455	2.706	3.841	5.412	6.635
2	.0201	.103	1.386	4.605	5.991	7.824	9.210
3	.115	.352	2.366	6.251	7.815	9.837	11.341
4	.297	.711	3.357	7.779	9.488	11.668	13.277
5	.554	1.145	4.351	9.236	11.070	13.388	15.086
6	.872	1.635	5.348	10.645	12.592	15.033	16.812
7	1.239	2.167	6.346	12.017	14.067	16.622	18.475
8	1.646	2.733	7.344	13.362	15.507	18.168	20.090
9	2.088	3.325	8.343	14.684	16.919	19.679	21.666
10	2.558	3.940	9.342	15.987	18.307	21.161	23.209
11	3.053	4.575	10.341	17.275	19.675	22.618	24.725
12	3.571	5.226	11.340	18.549	21.026	24.054	26.217
13	4.107	5.892	12.340	19.812	22.362	25.472	27.688
14	4.660	6.571	13.339	21 064	23.685	26.873	29.141
15	5.229	7.261	14.339	22.307	24.996	28.259	30.578
16	5.812	7.962	15.338	23.542	26.296	29.633	32.000
17	6.408	8.672	16.338	24.769	27.587	30.995	33.409
18	7.015	9 390	17.338	25.989	28.869	32.346	34.805
19	7.633	10.117	18.338	27.204	30.144	33.687	36.191
20	8.260	10.851	19.337	28.412	31.410	35.020	37.566
21	8.897	11.591	20.337	29.615	32.671	36.343	38.932
22	9.542	12.338	21.337	30.813	33.924	37.659	40.289
23	10.196	13.091	22.337	32.007	35 172	38.968	41.638
24	10.856	13.848	23.337	33.196	36.415	40.270	42.980
25	11.524	14.611	24.337	34.382	37.652	41.566	44.314
26	12.198	15.379	25.336	35.563	38.885	42.856	45.642
27	12.879	16.151	26.336	36.741	40.113	44.140	46.963
28	13.565	16.928	27.336	37.916	41.337	45.419	48.278
29	14.256	17.708	28.336	39.087	42.557	46.693	49.588
30	14.953	18.493	29.336	40.256	43.773	47.962	50.892

same for all values of n. When $\chi^2 = 0$, $P = 1.00$; when $\chi^2 = \infty$, $P = 0.00$. But for intermediate values the relationship varies with n.

In 1900 Karl Pearson defined the distribution function

[1] This table is reproduced here through the courtesy of R. A. Fisher and his publishers, Oliver and Boyd, of Edinburgh. The entries are taken from Table III of *Statistical Methods for Research Workers.*

of χ^2. The actual application of the χ^2 test is facilitated by prepared tables. Selected entries from these tabulations, for different values of n, are given in Table 143 on page 625 and in Appendix Table V.

For determining the significance of χ^2 beyond the range of this table, Fisher has given $\sqrt{2\chi^2} - \sqrt{2n-1}$, as a value which may be interpreted as a normal deviate. That is, the figure derived when stated values of χ^2 and n are inserted in the above expression is to be taken as a deviation from the mean of a normal distribution, expressed in units of the standard deviation of that distribution. The corresponding value of P is then derived from a table of areas under the normal curve.

The χ^2 test is applicable to a considerable variety of problems. Wherever, on rational grounds, a set of theoretical frequencies may be derived, for comparison with observed frequencies, this test is appropriate in judging of the significance of the discrepancy between the two sets of frequencies. In the following pages three applications of this test are exemplified.

THE CHI-SQUARE TEST OF GOODNESS OF FIT

When an ideal frequency curve, whether normal or of some other type, is fitted to an actual frequency distribution, theory and observation are being compared. A test of the concordance of the two (i.e., of goodness of fit) may be made by inspection, but such a test is obviously inadequate. Precision may be secured by employing the χ^2 test. The example in Table 144, relating to the distribution of telephone subscribers discussed in Chapter XIII, illustrates the procedure.

There are 15 classes in this distribution. Since the totals of columns (2) and (3) must be the same, one degree of freedom is lost. In arbitrarily making the mean and standard deviation of the theoretical distribution equal to those of the observed distribution, we lose two more degrees of freedom.

TABLE 144

Computation of χ^2 for Testing Goodness of Fit
Normal Curve of Error Fitted to Distribution of Telephone Subscribers

(1) *Class limits*	(2) *Observed frequency* f_0	(3) *Theoretical frequency* f	(4) $(f_0 - f)$	(5) $\frac{(f_0 - f)^2}{f}$
150 and less	10	13.14	− 3.14	.75
150–200	19	16.76	+ 2.24	.30
200–250	38	31.57	+ 6.43	1.31
250–300	50	53.02	− 3.02	.17
300–350	95	79.43	+ 15.57	3.05
350–400	85	106.10	− 21.10	4.20
400–450	115	126.41	− 11.41	1.03
450–500	132	134.31	− 2.31	.04
500–550	144	125.50	+ 18.50	2.73
550–600	116	106.51	+ 9.49	.85
600–650	79	81.85	− 2.85	.10
650–700	54	55.21	− 1.21	.03
700–750	31	33.19	− 2.19	.14
750–800	11	17.81	− 6.81	2.60
More than 800	16	14.19	+ 1.81	.23
	995	995.00	15 groups	$\chi^2 = 17.53$

In effect, three constants have been employed in the fitting process. The given value of χ^2, 17.53, is determined, therefore, with 12 degrees of freedom (the number of classes less 3).[1] From Table 143 we see that when $n = 12$ a value of χ^2 as great as 21.026, or greater, would occur purely as a result of chance in 5 out of 100 random samples, if the true value of χ^2 were zero. The value 17.53 is not excessively high. Observed discrepancies may be attributed to chance. The fit obtained with the normal curve is acceptable. Our results are not inconsistent with the hypothesis that the observations in question follow the normal law of error.

[1] This derivation of degrees of freedom, which differs from that originally employed by Pearson and exemplified in the first printing of this edition, gives a more accurate and logical test of significance than does the older procedure

In applying the Chi-square test it is not necessary to determine the exact probability corresponding to a stated value of χ^2. Our purpose, in general, is to ascertain whether observed results are or are not consistent with the hypothesis on which the fitting procedure is based. For this purpose we wish only to know whether the value of P corresponding to a given value of χ^2 falls below (or, much more rarely, above) certain critical values. As a conventional limit .05 is usually employed. If a value of χ^2 is such that P is below .05, the discrepancies between observed and theoretical values are, on this standard, considered too great to be attributed to chance. The hypothesis on the basis of which the theoretical frequencies have been determined is suspect, in such a case. If χ^2 is large enough to give values of P below .02 or .01, the inadequacy of the hypothesis is, of course, more strongly indicated.

R. A. Fisher points out that suspicion should attach to very low values of χ^2, which give values of P of .99 or thereabout. These values indicate a very close agreement between the hypothesis and the observed facts. Such close agreement may be due to chance, but there is strong probability that the hypothesis is at fault or, in mathematical terms, that the wrong function is being used. Coincidence of observed and theoretical values suggests the kind of agreement one obtains by fitting to n points a curve in the equation to which there are n constants. Any artificial forcing of agreement between hypothesis and observation of course invalidates the application of the Chi-square test.

In applying the Chi-square test it is convenient to use the conventional standards we have noted, as guides to the rejection or provisional acceptance of working hypotheses. It is unwise to use these standards arbitrarily, however. No single standard possesses significance in any absolute sense. The investigator in a given field of research will interpret the information such a test yields in the light of other knowledge relating to that field of experience, and with

due regard to the rational foundation of the hypotheses being tested.

One feature of Table 144 requires explanation. It will be noted that in the construction of this table the three classes at the lower end of the distribution have been lumped into one, and that the same thing has been done with the six classes at the upper end of the distribution (Cf. Tables 109 and 144). This is done to avoid the undue magnification of slight differences between the tails of the observed and theoretical distributions. When f, the theoretical frequency, is very small, a relatively slight absolute discrepancy between f_0 and f may serve to swell materially the value of χ^2. The lumping process is designed to prevent such a distortion. Since the selection of classes for combination rests on the personal judgment of the investigator, a subjective element is necessarily introduced here. However, the results of the test will not usually be much affected by minor variations in the combination of tail-end classes.[1]

The use of χ^2 in testing the fit of theoretical frequency curves is subject to another rather important limitation. In the computation of χ^2 no account is taken of the distribution of discrepancies between f_0 and f. Yet the manner in which these discrepancies are distributed may materially influence our judgment as to the goodness of a given fit. In such an example as that given in Table 144, the successive values of $f_0 - f$, counting from the lower limit of the x-scale, might be alternately positive and negative. Something approaching this alternation would be expected if chance factors alone accounted for the differences between observed and theoretical frequencies. But the differences might be distributed otherwise. All the values of $f_0 - f$ below the

[1] Considerations of the same sort suggest that a sample of reasonable size is needed for the valid application of the Chi-square test in curve fitting. Deming and Birge set 500 observations as the minimum required in a test of this type, if confidence is to be placed in the result. Yule and Kendall suggest a smaller number, but place emphasis on the need of an adequate number of theoretical observations (preferably not less than 10) in every class.

mode might be positive, while all the values above the mode might be negative. The cumulated discrepancies, as measured by χ^2, might be equal in the two cases, yet far more confidence would attach to a fit marked by alternations of plus and minus deviations than to one in which a series of positive deviations were bunched together on the scale, and negative discrepancies were correspondingly clustered. This limitation serves as a warning against purely mechanical use of the χ^2 test. Examination of the fit, and interpretion of χ^2 in the light of the actual distribution of discrepancies, are required in the application of this test.

THE CHI-SQUARE TEST OF INDEPENDENCE OF PRINCIPLES OF CLASSIFICATION[1]

A question that frequently arises in research has to do with the relation between two principles of classification. Thus, in studying commodity price movements during revivals after business depressions, we may divide all commodities into durable and non-durable classes. We may again divide them into those the prices of which precede the general average of commodity prices in the revival, and those that lag behind the general index. If the quality of durability has no relation to the timing of price recovery, the two principles of classification are independent. However, certain considerations relating to the character of demand for durable and non-durable goods lead us to believe that the durability of a good is related to the behavior of its market price during a period of business revival. It is possible to apply an objective test to determine whether these principles of classification are, in fact, related.

Observed frequencies are recorded in Table 145.[2]

[1] For a discussion of tests of independence and homogeneity see Chapter IV of *Statistical Methods for Research Workers*, by R. A. Fisher.

[2] Data from *The Behavior of Prices*, National Bureau of Economic Research, New York, 1927, with later additions.

TABLE 145

Observation

Two-fold Classification of 208 Commodities

Commodity group	*Observed frequencies*		
	Number preceding general index on price rise	*Number lagging behind general index on price rise*	*Total*
Durable goods	6	61	67
Non-durable goods	50	91	141
Total	56	152	208

The nature of the durability classification requires no explanation. The classification relating to the timing of price changes in business revival is based on the average behavior of each of the 208 commodities during 13 periods of business revival occurring between 1890 and 1936. The process of cross-classification gives four "cells" among which the 208 commodities are divided in the manner indicated in the table.

With the observed frequencies that constitute the entries in these four cells we may compare a set of theoretical frequencies, derived from the hypothesis that the durability of economic goods has no relation to the timing of price advances after business depressions. These expected frequencies are computed readily from the sub-totals. The 67 durable goods constitute 32.21 per cent of all the commodities, while the 141 non-durable goods constitute 67.79 per cent of the total. If durability has no relation to the timing of price advance, after depression, we should expect the 56 commodities that preceded the general index to be divided between durable and non-durable goods in this same proportion. That is, 32.21 per cent of the 56 commodities, or 18.04, should be durable, while 67.79 per cent of the 56, or 37.96, should be non-durable. Similarly, the 152 commodities lagging behind the general index in price revival should be divided between the durable and

non-durable categories in exactly the same way, 32.21 per cent in the durable class, 67.79 per cent in the non-durable. These expected frequencies, which conform to our hypothesis that the two principles of classification are independent, are given in Table 146.

TABLE 146

Expectation

Two-fold Classification of 208 Commodities

Commodity group	*Expected frequencies*		*Total*
	Number preceding general index on price rise	*Number lagging behind general index on price rise*	
Durable goods	18.04	48.96	67
Non-durable goods	37.96	103.04	141
Total	56.00	152.00	208

Chi-square is computed from the relation $\chi^2 = \Sigma \left\{ \frac{(f_0 - f)^2}{f} \right\}$, in the following manner:

$$\chi^2 = \frac{(6 - 18.04)^2}{18.04} + \frac{(50 - 37.96)^2}{37.96} + \frac{(61 - 48.96)^2}{48.96} + \frac{(91 - 103.04)^2}{103.04} = 16.222.$$

There are four components of Chi-square in this instance, but, as may readily be seen by reference to the table of expected frequencies, only one degree of freedom enters into its computation. The expected frequencies must yield the four group totals, 56, 152, 67, and 141. Accordingly, as soon as we fill one of the four cells set up by the process of cross-classification, the other three are definitely determined. Given 18.04, the expected number of durable goods preceding the general index in price revival, the entries in the other cells are fixed. Subtraction of 18.04 from 56 and 67 will fill two of them, and the filling of these determines the fourth.

For the interpretation of the given value of Chi-square we turn to Table 143, which is to be read with n, the number of degrees of freedom, equal to 1. If durability of economic goods has no relation to the timing of their price changes in revival, the two principles of classification employed above are independent and the true value of Chi-square is zero. Are the observed results consistent with this hypothesis? The entries in Table 143 indicate that if the true value were zero, a value as great as 3.841 would occur 5 times out of 100, as a result of chance fluctuations. A value as great as 6.635 would occur only 1 time out of 100. The present value of Chi-square, 16.222, represents a still smaller probability. The results are not consistent with the hypothesis we have set up. The differences between the observed and expected frequencies are too great to be attributed to the play of chance. Durability, and factors of demand and supply related thereto, appear to play a definite role in the timing of price advances in business revivals.

This test, it should be noted, does not define the relationship between durability of goods and the timing of price revival. It leads us to reject the hypothesis that durability has no bearing on the sequence of price advances in revival. If, on the basis of some other rational hypothesis, we could obtain a set of expected frequencies representing a definite relationship other than one of independence, this hypothesis could be tested in the same manner. From the present evidence, however, we may only conclude that the proportion of durable goods preceding the general price index on revival is smaller and the proportion of non-durable goods larger than would be expected if durability had no relation to the timing of price recovery after a business depression.

THE CHI-SQUARE TEST OF HOMOGENEITY

For each of eight major industrial groups we have records showing, for the year 1933, the number of corporations

reporting net incomes from their operations and the number reporting no net incomes (i.e., suffering deficits). The returns relate to a total of 492,649 corporations. Is this total a homogeneous whole, or does the division of corporations between those earning net incomes and those suffering deficits vary significantly from group to group? The records appear in Table 147.

TABLE 147

Comparison of Observed and Theoretical Frequencies

(Tabulations based on corporate income tax returns for 1933, by major industrial groups)

(1) Group	(2) Total number of returns	(3) Actual number of returns showing no net income (f_0)	(4) Theoretical (expected) number of returns showing no net income (f)	(5) $f_0 - f$	(6) Actual number of returns showing net income (f_0')	(7) Theoretical (expected) number showing net income (f')	(8) $f_0' - f'$
Agriculture and related industries	10,490	7,818	7,150	+ 668	2,672	3,340	− 668
Mining and quarrying	17,147	8,866	11,688	− 2,822	8,281	5,459	+ 2,822
Manufacturing	93,833	62,295	63,958	− 1,663	31,538	29,875	+ 1,663
Construction	18,234	14,112	12,428	+ 1,684	4,122	5,806	− 1,684
Transportation and other public utilities	24,302	14,349	16,564	− 2,215	9,953	7,738	+ 2,215
Trade	137,858	93,621	93,965	− 344	44,237	43,893	+ 344
Service	47,843	35,419	32,610	+ 2,809	12,424	15,233	− 2,809
Finance	142,942	99,314	97,431	+ 1,883	43,628	45,511	− 1,883
Total	492,649	335,794	335,794		156,855	156,855	
Per cent	100.000	68.161			31.839		

The observed frequencies are, of course, the actual returns given in cols. (3) and (6) of Table 147. Sets of theoretical or expected frequencies may be set up, for comparison with these, on the assumption that all corporations in the United States constituted a homogeneous population as regards the likelihood of suffering a deficit or earning a net income in 1933. On this assumption we may say that the probability of failing to earn a net profit was, for all the elements of this assumed homogeneous population, $\frac{335,794}{492,649}$ or .68161. Conversely, the probability of earning a net income was $\frac{156,855}{492,649}$

or .31839. If these are the true probabilities for all elements of the population, we may determine theoretical frequencies for each industrial group by applying these ratios to the total number of corporations in that group. On the assumption made we should find, in all groups, the same proportionate division between corporations earning net incomes and those suffering deficits, except for modifications due to fluctuations of sampling. The expected frequencies appear in columns (4) and (7). If the hypothesis of homogeneity is valid, these are the true frequencies for the several groups. Differences between these and the observed frequencies reflect the play of chance alone.

The calculation of χ^2, measuring the degree of discrepancy between the observed and theoretical frequencies, is shown in the next table. It will be noted that there are two components of χ^2, one derived from the comparison of actual and theoretical frequencies among the corporations suffering deficits, the other from similar data for corporations earning net incomes. The symbols f_0 and f are used with prime marks for the second group, to avoid ambiguity, but both elements enter into the final measurement desired.

(1)	(2)	(3)	(4)	(5)	(6)
Group	$f_0 - f$	$(f_0 - f)^2$	$\frac{(f_0 - f)^2}{f}$	$f_0' - f'$	$\frac{(f_0' - f')^2}{f'}$
Agriculture and related industries	+ 668	446,224	62.4090	− 668	133.6000
Mining and quarrying	− 2,822	7,963,684	681.3555	+ 2,822	1,458.8174
Manufacturing	− 1,663	2,765,569	43.2404	+ 1,663	92.5713
Construction	+ 1,684	2,835,856	228.1828	− 1,684	488.4354
Transportation and other public utilities	− 2,215	4,906,225	296.1980	+ 2,215	634.0430
Trade	− 344	118,336	1.2594	+ 344	2.6960
Service	+ 2,809	7,890,481	241.9650	− 2,809	517.9860
Finance	+ 1,883	3,545,689	36.3918	− 1,883	77.9084
Total			1,591.0019		3,406.0575

The required value of χ^2 is equal to the sum of the entries in columns (4) and (6) of the above table. That is, χ^2 equals 1,591.0019 + 3,406.0575, or 4,997.0594. This is derived from 16 individual entries, 2 for each of the 8 classes. There are, however, only 7 degrees of freedom in its derivation. For as soon as 7 of the 16 cells for which frequencies are to be determined have been filled in, the other 9 are fixed. (The sum of the number of returns showing no net incomes and those showing net incomes must equal the actual number of returns for each class, and the grand total of theoretical frequencies must equal total actual frequencies.) Since the 1% value of χ^2, for $n = 7$, is only 18.475, it is clear that the discrepancy is too great to be attributed to chance. The results are not consistent with the hypothesis of homogeneity. We are not justified in assuming that the forces affecting the profitability of corporate operations in 1933 were the same, among the eight major industrial groups here represented.

The various procedures discussed in this chapter give some indication of the variety and power of the methods available for use in interpreting and appraising the results of statistical research. Each one involves, in some form, the testing of hypotheses against evidence yielded by the study of samples. It should be emphasized that the formal procedures described in the preceding pages are employed at a rather late stage in actual research work. The experiment will have been planned, the field work done, hypotheses framed before the tests here discussed can be applied. These various steps must, of course, be coördinated. The data must be gathered with reference to the hypotheses to be tested and to the analytical methods to be employed. Acquaintance with appropriate techniques is one prerequisite of intelligent planning of research in which quantitative data are utilized. Familiarity with the characteristics and limitations of the available materials, and clear definition of the questions at issue, are equally important elements.

REFERENCES

Camp, B. H., *The Mathematical Part of Elementary Statistics.* Part II, Chap. 4.

Deming, W. E. and Birge, R. T., "On the Statistical Theory of Errors." Reprint from *Reviews of Modern Physics*, Vol. 6, July, 1934. (Available from Graduate School, U. S. Department of Agriculture.)

Fisher, R. A., *Statistical Methods for Research Workers*, Chaps. 4–6.

Fisher, R. A., *The Design of Experiments.*

Neyman, J., *Lectures and Conferences on Mathematical Statistics*, with the editorial assistance of W. E. Deming. Graduate School, U. S. Department of Agriculture.

Shewhart, W. A., *Economic Control of Quality of Manufactured Product*, Chaps. 13–16.

Snedecor, G. W., *Statistical Methods*, Chaps. 1, 3, 9.

"Student," "The Probable Error of the Mean," *Biometrika*, Vol. 6, 1908.

Tippett, L. H. C., *The Methods of Statistics*, Chaps. 4, 5, 10.

Wilks, S. S., *Theory of Statistical Inference, 1936–1937.*

Yule, G. U. and Kendall, M. G., *An Introduction to the Theory of Statistics*, Chaps. 22, 23.

For discussion of sampling problems in time series analysis see:

Roos, Charles F., "The Correlation and Analysis of Time Series," *Econometrica*, Vol. 4, no. 4, 368–381.

Schultz, Henry, "The Standard Error of a Forecast from a Curve," *Journal of the American Statistical Association*, Vol. 25, no. 170, 139–185.

Working, Holbrook and Hotelling, Harold, "The Application of the Theory of Error to the Interpretation of Trends," *Proceedings of the American Statistical Association*, March, 1929.

There is an extensive body of recent literature on the subjects briefly discussed in Chapter XVIII. No complete list of these publications may be given here. In addition to the general references given above we may note the following:

Fisher, R. A., "The Mathematical Distributions Used in the Common Tests of Significance," *Econometrica*, Vol. 3, no. 4, 353–365.

Hotelling, Harold, "The Generalization of Student's Ratio," *Annals of Mathematical Statistics*, Vol. 2, 1931, 360–378.

Pearson, E. S., "Sampling Problems in Industry," *Journal of the Royal Statistical Society*, Vol. 1, no. 2, 1934, *Supplement.*

Rider, P. R., "A Survey of the Theory of Small Samples," *Annals of Mathematics*, Vol. 31, 1930.

Wilks, S. S., "Test Criteria for Statistical Hypotheses Involving Several Variables," *Journal of the American Statistical Association*, Vol. 30, 1935, 549–560.

APPENDIX A

THE METHOD OF LEAST SQUARES AS APPLIED TO CERTAIN STATISTICAL PROBLEMS

The method of least squares in the case of a single unknown quantity is merely a procedure for obtaining the most probable value of that quantity from a number of separate observations. The most probable value is that for which the sum of the squares of the deviations (or residuals) is a minimum. This is the arithmetic mean of the observations.

Where the measurements or observations do not relate directly to a single unknown quantity, but to *functions* of a number of unknown quantities, the problem is somewhat different. In the first case mentioned each observation is in the form of a single magnitude. In the present case each observation is in the form of an *observation equation* in which the observed values of the variables, as found in combination, are entered. The unknown quantities are the constants which define the functional relationship between the variables in question. Our problem is that of finding the most probable values of these constants, the true values being unknown.

As in the simpler case the most probable values are those for which the sum of the squares of the residuals is a minimum. In this case, however, the residuals are deviations, not from a single magnitude, as in the case of the arithmetic mean, but from the curve which describes the most probable functional relationship. The residuals are the differences between the computed and the actual values of the dependent variable.

DERIVATION OF THE NORMAL EQUATIONS

Representing by Y an observed value of the dependent variable, by Y_c the corresponding computed value, by v the residual, or difference between Y and Y_c, and by W_1, W_2, W_3, and W_4 different independent variables (or different functions of a single independent variable), we may write

$$\begin{aligned} Y_c &= f(W_1, W_2, W_3, W_4) \\ v &= Y_c - Y \\ &= f(W_1, W_2, W_3, W_4) - Y \\ \Sigma(v^2) &= \Sigma[f(W_1, W_2, W_3, W_4) - Y]^2. \end{aligned}$$

If the function in a particular case is of the type

$$Y_c = aW_1 + bW_2 + cW_3 + dW_4$$

we have

$$\Sigma(v^2) = \Sigma[(aW_1 + bW_2 + cW_3 + dW_4) - Y]^2.$$

Our problem is that of determining the most probable values of the constants that define the function. These constants are represented, in the present case, by a, b, c, and d. (The W's, it should be noted, refer to quantities which are known, once the observation equations are given. In the usual case the W's are different functions of a single variable, but this is not essential.) On the assumption that the errors of observation are distributed in accordance with the normal law of error, it may be demonstrated that the most probable values of a, b, c, and d, in the above equation, are those which render $\Sigma(v^2)$ a minimum; i.e.,

$$\Sigma[(aW_1 + bW_2 + cW_3 + dW_4) - Y]^2 = \text{a minimum.} \qquad (A)$$

The normal equations necessary for the solution may be obtained by equating to zero the partial derivatives of the above expression with respect to the unknowns, a, b, c, and d. That is, we first differentiate the above function with respect to a, holding b, c, and d constant, then with respect to b, holding a, c, and d constant, then with respect

to c, holding a, b, and d constant, then with respect to d, holding a, b, and c constant. Carrying through this operation with respect to a, we have

$$\frac{\partial}{\partial a}\Sigma[(aW_1 + bW_2 + cW_3 + dW_4) - Y]^2 = 0$$

or

$$\text{I} \quad \Sigma W_1[(aW_1 + bW_2 + cW_3 + dW_4) - Y] = 0.$$

Differentiating equation (A) now with respect to b, we have

$$\frac{\partial}{\partial b}\Sigma[(aW_1 + bW_2 + cW_3 + dW_4) - Y]^2 = 0$$

or

$$\text{II} \quad \Sigma W_2[(aW_1 + bW_2 + cW_3 + dW_4) - Y] = 0.$$

Differentiating equation (A) with respect to c,

$$\frac{\partial}{\partial c}\Sigma[(aW_1 + bW_2 + cW_3 + dW_4) - Y]^2 = 0$$

or

$$\text{III} \quad \Sigma W_3[(aW_1 + bW_2 + cW_3 + dW_4) - Y] = 0.$$

Differentiating equation (A) with respect to d,

$$\frac{\partial}{\partial d}\Sigma[(aW_1 + bW_2 + cW_3 + dW_4) - Y]^2 = 0$$

or

$$\text{IV} \quad \Sigma W_4[(aW_1 + bW_2 + cW_3 + dW_4) - Y] = 0.$$

The most probable values of the quantities a, b, c, and d are secured by solving simultaneously the four normal equations thus obtained (numbered above I, II, III, IV).

FORMATION OF THE NORMAL EQUATIONS

When the observation equations are all of the first degree (i.e., of the first degree with respect to the unknown quantities, a, b, c, etc.) the normal equations may be secured by the following process:

1. Write the equation which describes the assumed relationship. The observation equations are derived by substituting in this equation the observed values of the variables, as found in combination.

2. Multiply each observation equation by the coefficient of the first unknown in that equation; the sum of the resulting equations constitutes the first normal equation.

3. Multiply each observation equation by the coefficient of the second unknown in that equation; the sum of the resulting equations constitutes the second normal equation.

Continue this process until normal equations equal in number to the unknown quantities are obtained.

The actual process of forming the normal equations in curve fitting may be simplified, and the writing out of the separate observation equations avoided, as was demonstrated in earlier sections. The following may be laid down as general rules for the formation of the desired normal equations:

1. Write the equation of the curve to be fitted. For the purpose of this explanation we may employ the general form

$$Y = aW_1 + bW_2 + cW_3 + dW_4 + \ldots \qquad (1)$$

where Y represents the dependent variable, a, b, c, d, . . . represent the constants in the equation (the unknown quantities in the present instance) and W_1, W_2, W_3, W_4, . . . represent the coefficients of these unknowns. It is assumed that these coefficients represent variables, and that term is used with reference to them. Call this equation (1).

2. Multiply each term in equation (1) by the coefficient of the first unknown in (1) (i.e., by W_1) and place the summation sign, Σ, before each variable. This is the first normal equation (I).

3. Multiply each term in equation (1) by the coefficient of the second unknown (i.e., by W_2) and place the summation sign before each variable. This is the second normal equation (II).

4. Multiply each term in equation (1) by the coefficient of the third unknown (i.e., by W_3) and place the summation sign before each variable. This is the third normal equation (III).

5. Multiply each term in equation (1) by the coefficient of the fourth unknown (i.e., by W_4) and place the summation sign before each variable. This is the fourth normal equation (IV).

The process may be continued until normal equations equal in number to the unknown quantities are obtained.[1]

A STANDARD SET OF NORMAL EQUATIONS

As a set of generalized normal equations secured by the above process and applying to any equation which can be put in the form

$$Y = aW_1 + bW_2 + cW_3 + dW_4 + \ldots,$$

we have

$$\begin{aligned}
\text{I}\quad & \Sigma(W_1Y) \\
& = a\Sigma(W_1^2) + b\Sigma(W_1W_2) + c\Sigma(W_1W_3) + d\Sigma(W_1W_4) + \ldots \\
\text{II}\quad & \Sigma(W_2Y) \\
& = a\Sigma(W_1W_2) + b\Sigma(W_2^2) + c\Sigma(W_2W_3) + d\Sigma(W_2W_4) + \ldots \\
\text{III}\quad & \Sigma(W_3Y) \\
& = a\Sigma(W_1W_3) + b\Sigma(W_2W_3) + c\Sigma(W_3^2) + d\Sigma(W_3W_4) + \ldots \\
\text{IV}\quad & \Sigma(W_4Y) \\
& = a\Sigma(W_1W_4) + b\Sigma(W_2W_4) + c\Sigma(W_3W_4) + d\Sigma(W_4^2) + \ldots
\end{aligned}$$

By substituting for W_1, W_2, W_3, W_4, etc., the particular functions employed in a given case, these equations may be readily adapted to any type of curve in the fitting of which the method of least squares is applicable. Thus in fitting a curve represented by the equation

$$Y = a + bX + cX^2 + dX^3$$

substitutions in the standard normal equations given above are based upon the following relations:

$$\begin{aligned}
W_1 &= 1 \\
W_2 &= X \\
W_3 &= X^2 \\
W_4 &= X^3.
\end{aligned}$$

The changes to be made in the normal equations are obvious. $\Sigma(W_1Y)$ becomes $\Sigma(Y)$; $\Sigma(W_1^2)$ is equivalent to $\Sigma(1^2)$, which is equal to N, the total number of observations.

[1] These rules represent an adaptation of a similar series formulated by Raymond Pearl in *Medical Biometry and Statistics*, 341.

The first normal equation becomes

$$\Sigma(Y) = Na + b\Sigma(X) + c\Sigma(X^2) + d\Sigma(X^3).$$

The other normal equations are modified correspondingly.

In the example just given, the coefficients are all different functions of a single independent variable, X. It is not, of course, essential to the method of least squares that this be so. The coefficients, W_1, W_2, W_3, etc., may represent a number of independent variables, as in the case of multiple correlation.

The limitations to the method of least squares must be borne in mind in making use of it. This method, in its direct application, is limited to cases in which the equation to the curve to be fitted is linear in the constants, i.e., the observation equations must all be linear as regards the unknown values, a, b, c, etc. (This does not mean, of course, that the equation to the fitted curve must be linear.) As an example of this limitation, we may cite a curve having as equation $y = ab^{c^x}$, which cannot be fitted directly by the method of least squares. If the observation equations are non-linear they may be reduced to the linear form in many instances by the use of logarithms, and the method of least squares then employed.

DERIVATION OF THE FORMULA FOR THE STANDARD ERROR OF ESTIMATE

It has been pointed out in the body of the text that the standard error of estimate may be derived as a by-product of the method of least squares. A more complete demonstration of this process may be given at this point.

When the partial derivative of the expression

$$\Sigma[(aW_1 + bW_2 + cW_3 + dW_4) - Y]^2 = \text{a minimum}$$

is equated to zero, with respect to the first unknown, a, we have

$$\Sigma W_1[(aW_1 + bW_2 + cW_3 + dW_4) - Y] = 0.$$

Since

$$aW_1 + bW_2 + cW_3 + dW_4 - Y = v,$$

we have as a necessary condition of fitting

$$\Sigma(vW_1) = 0.$$

When the partial derivative of the same expression with respect to b is equated to zero, we have

$$\Sigma W_2[(aW_1 + bW_2 + cW_3 + dW_4) - Y] = 0$$

or, making the same substitution as in the preceding case,

$$\Sigma(vW_2) = 0.$$

Repeating the operation with respect to c and d, we may show that

$$\Sigma(vW_3) = 0$$

and

$$\Sigma(vW_4) = 0.$$

In summary: When the method of least squares is employed in determining the most probable values of certain unknown quantities, having as known coefficients the quantities W_1, W_2, W_3, W_4, the following relations hold as a necessary condition of the least squares method:

$$\begin{aligned}\Sigma(vW_1) &= 0\\ \Sigma(vW_2) &= 0\\ \Sigma(vW_3) &= 0\\ \Sigma(vW_4) &= 0.\end{aligned}$$

A knowledge of these relationships gives us a method of securing readily the value $\Sigma(v^2)$ and the standard error of estimate. Assume that, by the method of least squares, we have determined the constants in an equation of the type

$$Y_c = aW_1 + bW_2 + cW_3 + dW_4.$$

For each residual we have the relation

$$v = aW_1 + bW_2 + cW_3 + dW_4 - Y. \qquad (1)$$

Multiplying throughout by v, and summing, we have

$$\Sigma(v^2) = a\Sigma(vW_1) + b\Sigma(vW_2) + c\Sigma(vW_3) + d\Sigma(vW_4) - \Sigma(Yv). \quad (2)$$

But

$$\begin{aligned} \Sigma(vW_1) &= 0 \\ \Sigma(vW_2) &= 0 \\ \Sigma(vW_3) &= 0 \\ \Sigma(vW_4) &= 0 \end{aligned}$$

therefore,

$$\Sigma(v^2) = -\Sigma(Yv). \quad (3)$$

Multiplying each equation (1) throughout by Y, and adding, we have

$$\Sigma(Yv) = a\Sigma(W_1Y) + b\Sigma(W_2Y) + c\Sigma(W_3Y) + d\Sigma(W_4Y) - \Sigma(Y^2). \quad (4)$$

Substituting in (3) the equivalent of $\Sigma(Yv)$, we have

$$\Sigma(v^2) = \Sigma(Y)^2 - a\Sigma(W_1Y) - b\Sigma(W_2Y) - c\Sigma(W_3Y) - d\Sigma(W_4Y). \quad (5)$$

This gives us a method of obtaining the value $\Sigma(v^2)$ without computing the separate residuals, a method that is applicable whenever the equation of the curve to be fitted is of the form, or may be reduced by the use of logarithms, reciprocals, or other manipulation to the form

$$Y = aW_1 + bW_2 + cW_3 + dW_4.$$

In applying this to a particular case it is necessary only to replace W_1, W_2, W_3, W_4, etc., by the functions that actually appear as coefficients of the unknown quantities in the original equation. Thus in fitting a curve the equation to which is

$$Y = a + bX + cX^2 + dX^3,$$

we find, as noted above, that

$$\begin{aligned} W_1 &= 1 \\ W_2 &= X \\ W_3 &= X^2 \\ W_4 &= X^3. \end{aligned}$$

Making these substitutions in equation (5) above, we have

$$\Sigma(v^2) = \Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY) - c\Sigma(X^2Y) - d\Sigma(X^3Y). \quad (6)$$

The standard error, S_y, is derived from the equation

$$S_y^2 = \frac{\Sigma(d^2)}{N}^*$$

where d is used to represent a deviation from a fitted curve. The deviation, d, then, is but another term for the residual v. Accordingly, as a general expression for the standard error of Y, with W_1, W_2, W_3, and W_4 as independent variables, we have

$$S_y^2 = \frac{\Sigma Y^2 - a\Sigma(W_1Y) - b\Sigma(W_2Y) - c\Sigma(W_3Y) - d\Sigma(W_4Y)}{N}. \quad (7)$$

As in the previous case, this may be applied to a particular problem by replacing W_1, W_2, W_3, W_4, etc., by the actual coefficients of the unknown quantities.

DERIVATION OF THE FORMULA FOR THE INDEX OF CORRELATION

We have adopted as an index of the degree of correlation between two variables the measure ρ (rho), derived from the equation

$$\rho^2_{yx} = 1 - \frac{S_y^2}{\sigma_y^2} \quad (8)$$

assuming a single dependent variable, Y, and a single independent variable, X. With a single dependent variable, Y, and a number of independent variables, W_1, W_2, W_3, W_4, the expression might be written

$$\rho^2_{y \cdot w_1 w_2 w_3 w_4} = 1 - \frac{S_y^2}{\sigma_y^2}. \quad (9)$$

* Since our object is to measure the actual "scatter" about the fitted curve, the formula $\frac{\Sigma(d^2)}{N}$ is used, rather than the formula $\frac{\Sigma(d^2)}{N - N_c}$ (where N represents the number of observations and N_c the number of constants in the equation to the fitted curve). The second formula would be used, in accordance with the theory of least squares, if we were seeking to determine the mean square error of an observation or of an observational equation.

Corresponding changes would be made in the subscripts for other changes in the symbols employed. The expression above is equivalent to

$$\rho^2_{y \cdot w_1 w_2 w_3 w_4} = 1 - \frac{\Sigma(d^2)}{\Sigma(y^2)}$$

where y represents a deviation from an origin at the mean of the Y's. But

$$\frac{\Sigma(y^2)}{N} = \frac{\Sigma(Y^2)}{N} - c_y^2$$

where Y represents the original values of the Y-variable and c_y represents the difference between the original origin and the mean of the Y's. (The symbols c_y, and c_x should not be confused with c, one of the constants in the equation to the fitted curve.)

Accordingly, we have

$$\rho^2_{y \cdot w_1 w_2 w_3 w_4} = 1 - \frac{\Sigma(d^2)}{\Sigma(Y^2) - Nc_y^2}. \tag{10}$$

But we have secured an expression for $\Sigma(v^2)$ [the equivalent of $\Sigma(d^2)$] which holds in the case of a curve fitted by the method of least squares. Substituting the equivalent of $\Sigma(d^2)$ in the above equation, and simplifying, we have, as a general formula for the index of correlation

$$\rho^2_{y \cdot w_1 w_2 w_3 w_4 \ldots} = \tag{11}$$
$$\frac{a\Sigma(W_1Y) + b\Sigma(W_2Y) + c\Sigma(W_3Y) + d\Sigma(W_4Y) + \ldots - Nc_y^2}{\Sigma(Y^2) - Nc_y^2}$$

This may be applied to a specific case by replacing W_1, W_2, W_3, W_4, etc., in the above formula by the functions which appear as coefficients of the unknown quantities in the original equation. When all these are functions of a single independent variable, as in the usual case, the index of correlation would be represented by the symbol ρ_{yx}.

CERTAIN SPECIAL CASES

In the case of multiple correlation, where the symbols X_1, X_2, X_3, X_4, etc., are used to represent all the variables,

whether considered dependent or independent, the symbol R is employed for the measure of correlation and numerical subscripts utilized as described in the body of the text.

In the case of a straight line relationship between two variables, ρ is replaced by the symbol r, which represents the ordinary coefficient of correlation. As the general equation for r we have

$$r^2 = \frac{a\Sigma(Y) + b\Sigma(XY) - Nc_y^2}{\Sigma(Y^2) - Nc_y^2}.$$

There are two special cases in which this formula may be simplified. If the origin be at the mean of the X's, we have

$$a = c_y = \frac{\Sigma Y}{N}$$

$$a^2 = c_y^2 = \frac{a\Sigma Y}{N}$$

$$Nc_y^2 = a\Sigma Y$$

and the formula for r reduces to

$$r^2 = \frac{b\Sigma(xY)}{\Sigma(Y^2) - Nc_y^2}.$$

If the origin be at the mean of the Y's (it is not essential that it be also at the mean of the X's)

$$\Sigma(y) = 0, \text{ and } c_y = 0$$

and the formula for the coefficient of correlation becomes

$$r^2 = \frac{b\Sigma(Xy)}{\Sigma(y^2)}.$$

In this latter case the general formula for ρ may also be simplified by the elimination of the terms $a\Sigma(y)$ and Nc_y^2.

CHECKS ON THE FORMATION OF THE NORMAL EQUATIONS

There are so many possibilities of arithmetical error in the formation and solution of a set of normal equations that checks should be employed wherever possible. A convenient check on the calculations leading to the normal

equations is afforded by the introduction in each observation equation of an additional term, s, equal to the sum of all the known quantities in that equation. Thus, in the following system of observation equations, formed in fitting a line to the points 1, 3; 2, 4; 3, 6; 4, 5; 5, 10; 6, 9; 7, 10; 8, 12; 9, 11, the values of s are as indicated:

	s
$3 = a + 1b$	5
$4 = a + 2b$	7
$6 = a + 3b$	10
$5 = a + 4b$	10
$10 = a + 5b$	16
$9 = a + 6b$	16
$10 = a + 7b$	18
$12 = a + 8b$	21
$11 = a + 9b$	21.

(The coefficient of a in each case is 1, and this is added to the other known quantities.)

In fitting a curve described by the type equation

$$Y = aW_1 + bW_2 + cW_3 + dW_4$$

the following relations prevail between s and the other quantities computed. For each observation equation,

$$Y + W_1 + W_2 + W_3 + W_4 = s.$$

For the normal equations,

$$\Sigma(W_1Y) + \Sigma(W_1^2) + \Sigma(W_1W_2) + \Sigma(W_1W_3) + \Sigma(W_1W_4) = \Sigma(W_1s)$$
$$\Sigma(W_2Y) + \Sigma(W_1W_2) + \Sigma(W_2^2) + \Sigma(W_2W_3) + \Sigma(W_2W_4) = \Sigma(W_2s)$$
$$\Sigma(W_3Y) + \Sigma(W_1W_3) + \Sigma(W_2W_3) + \Sigma(W_3^2) + \Sigma(W_3W_4) = \Sigma(W_3s)$$
$$\Sigma(W_4Y) + \Sigma(W_1W_4) + \Sigma(W_2W_4) + \Sigma(W_3W_4) + \Sigma(W_4^2) = \Sigma(W_4s)$$

This form is capable of application to any specific problem. In each case the s-equations are formed in precisely the same way as the corresponding normal equations.

In applying these checks several additional columns are needed in the working tables, but the extra trouble is more than compensated by the opportunity to check the

work at each stage. The application is illustrated in the following working table, showing the calculations involved in fitting a second degree curve of the form

$$Y = a + bX + cX^2$$

to the nine points 1, 2; 2, 6; 3, 7; 4, 8; 5, 10; 6, 11; 7, 11; 8, 10; 9, 9.

TABLE A

Illustrating the Use of Checks on the Formation of Normal Equations

Y	X	X^2	XY	X^2Y	s	Xs	X^2s
2	1	1	2	2	5	5	5
6	2	4	12	24	13	26	52
7	3	9	21	63	20	60	180
8	4	16	32	128	29	116	464
10	5	25	50	250	41	205	1,025
11	6	36	66	396	54	324	1,944
11	7	49	77	539	68	476	3,332
10	8	64	80	640	83	664	5,312
9	9	81	81	729	100	900	8,100
74	45	285	421	2,771	413	2,776	20,414

(Columns for X^3 and X^4 are omitted, as the values $\Sigma(X^3)$ and $\Sigma(X^4)$ may be derived from prepared tables.)

Each of the values in the column headed s is secured from the corresponding observation equation. Thus, from the first observation equation

$$2 = 1a + 1b + 1c,$$

we have 5 as the value of s (2, plus the coefficients of the three constants). These values of s are secured readily from the table by adding the figures in the columns headed Y, X, and X^2, plus 1, the coefficient of the constant term a.

Adding the various columns, the arithmetic work is verified by the following checks:

$$\Sigma(Y) + N + \Sigma(X) + \Sigma(X^2) = \Sigma(s)$$
$$74 + 9 + 45 + 285 = 413$$
$$\Sigma(XY) + \Sigma(X) + \Sigma(X^2) + \Sigma(X^3) = \Sigma(Xs)$$

$$421 + 45 + 285 + 2{,}025 = 2{,}776$$
$$\Sigma(X^2Y) + \Sigma(X^2) + \Sigma(X^3) + \Sigma(X^4) = \Sigma(X^2s)$$
$$2{,}771 + 285 + 2{,}025 + 15{,}333 = 20{,}414.$$

Further uses of a check of this kind are explained below, in discussing the solution of the normal equations.

OTHER TESTS

The possibility of checking the calculations in other ways has been suggested in the preceding sections. Thus, where the coefficients of the constants in the equation to the fitted curve are represented by W_1, W_2, W_3, W_4, we know that

$$\Sigma(vW_1) = 0$$
$$\Sigma(vW_2) = 0$$
$$\Sigma(vW_3) = 0$$
$$\Sigma(vW_4) = 0.$$

If a curve of the type

$$Y = a + bX + cX^2 + dX^3$$

has been fitted, this means that

$$\Sigma(v) = 0$$
$$\Sigma(vX) = 0$$
$$\Sigma(vX^2) = 0$$
$$\Sigma(vX^3) = 0.$$

The accuracy of the work may be tested by checking these relations.

Finally, we may test the accuracy of the work by computing the standard error of estimate in two different ways. We may compute the separate residuals by taking the difference between computed and actual values of the dependent variable, and from these values determine S. This may be compared with the results secured by applying the general formula for the standard error, as derived above. In the fitting of the second degree curve, the data of which

were used to illustrate the method of checking the normal equations, the equation derived was

$$Y = -.92860 + 3.52316X - .267316X^2.$$

From the residuals separately computed, we have

$$S_y = .4941.$$

From the formula

$$S_y^2 = \frac{\Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY) - c\Sigma(X^2Y)}{N},$$

we have

$$S_y = .4947.$$

This constitutes a final check upon the accuracy of the calculations.

SIMPLIFICATION OF NORMAL EQUATIONS IN A MULTIPLE CORRELATION PROBLEM[1]

In the discussion of multiple correlation procedure in Chapter XVI the normal equations as first derived in the form

$$\begin{aligned}
\text{I}\quad & \Sigma(X_1) = Na + b_{12.34}\Sigma(X_2) + b_{13.24}\Sigma(X_3) + b_{14.23}\Sigma(X_4) \\
\text{II}\quad & \Sigma(X_1X_2) = a\Sigma(X_2) + b_{12.34}\Sigma(X_2^2) + b_{13.24}\Sigma(X_2X_3) \\
& \qquad + b_{14.23}\Sigma(X_2X_4) \\
\text{III}\quad & \Sigma(X_1X_3) = a\Sigma(X_3) + b_{12.34}\Sigma(X_2X_3) + b_{13.24}\Sigma(X_3^2) \\
& \qquad + b_{14.23}\Sigma(X_3X_4) \\
\text{IV}\quad & \Sigma(X_1X_4) = a\Sigma(X_4) + b_{12.34}\Sigma(X_2X_4) + b_{13.24}\Sigma(X_3X_4) \\
& \qquad + b_{14.23}\Sigma(X_4^2)
\end{aligned}$$

were reduced in number and modified to facilitate their solution. Details of the method are here given.

Letting A_1, A_2, A_3, and A_4 represent the arithmetic means of the several variables, and x_1, x_2, x_3, and x_4 represent deviations from the means, we may replace the variables

[1] Adapted from H. R. Tolley and M. J. B. Ezekiel, "A Method of Handling Multiple Correlation Problems," *Journal of the American Statistical Association*, Vol. 18, 993–1003.

X_1, X_2, X_3, and X_4 by their equivalents $x_1 + A_1$, $x_2 + A_2$, $x_3 + A_3$, $x_4 + A_4$. The normal equations now become:

$$\text{I}\quad \Sigma(x_1 + A_1) = Na + \Sigma(x_2 + A_2) \cdot b_{12.34} + \Sigma(x_3 + A_3) \cdot b_{13.24} + \Sigma(x_4 + A_4) \cdot b_{14.23}$$

$$\text{II}\quad \Sigma[(x_1 + A_1)(x_2 + A_2)] = \Sigma[(x_2 + A_2) \cdot a + \Sigma(x_2 + A_2)^2] \cdot b_{12.34} + \Sigma[(x_2 + A_2)(x_3 + A_3)] \cdot b_{13.24} + \Sigma(x_2 + A_2)(x_4 + A_4) \cdot b_{14.23}$$

$$\text{III}\quad \Sigma[(x_1 + A_1)(x_3 + A_3)] = \Sigma(x_3 + A_3) \cdot a + \Sigma[(x_3 + A_3)(x_2 + A_2)] \cdot b_{12.34} + \Sigma(x_3 + A_3)^2 \cdot b_{13.24} + \Sigma[(x_3 + A_3)(x_4 + A_4)] \cdot b_{14.23}$$

$$\text{IV}\quad \Sigma[(x_1 + A_1)(x_4 + A_4)] = \Sigma(x_4 + A_4) \cdot a + \Sigma[(x_4 + A_4)(x_2 + A_2)] \cdot b_{12.34} + \Sigma[(x_4 + A_4)(x_3 + A_3)] \cdot b_{13.24} + \Sigma(x_4 + A_4)^2 \cdot b_{14.23}.$$

Since $\Sigma(x_1 + A_1) = \Sigma x_1 + NA_1$, and since $\Sigma x_1 = 0$, $\Sigma(x_1 + A_1)$ and all similar expressions may be replaced by NA_1, NA_2, etc.

If we expand $\Sigma(x_2 + A_2)^2$ to $\Sigma(x_2^2 + 2A_2x_2 + A_2^2)$, the middle term drops out, because $\Sigma x_2 = 0$, and the expression may be written $\Sigma x_2^2 + NA_2^2$. The sums of all similar squares may be put in similar form.

The product sum $\Sigma(x_1 + A_1)(x_2 + A_2) = \Sigma(x_1x_2 + A_1x_2 + A_2x_1 + A_1A_2) = \Sigma x_1x_2 + NA_1A_2$ since $\Sigma x_1 = 0$ and $\Sigma x_2 = 0$. Product sums of the same type may be similarly modified. The normal equations now take the form:

$$\text{I}\quad NA_1 = Na + NA_2b_{12.34} + NA_3b_{13.24} + NA_4b_{14.23}$$

$$\text{II}\quad \Sigma(x_1x_2) + NA_1A_2 = NA_2a + [\Sigma(x_2)^2 + NA_2^2]b_{12.34} + [\Sigma(x_2x_3) + NA_2A_3]b_{13.24} + [\Sigma(x_2x_4) + NA_2A_4]b_{14.23}$$

$$\text{III}\quad \Sigma(x_1x_3) + NA_1A_3 = NA_3a + [\Sigma(x_2x_3) + NA_2A_3]b_{12.34} + [\Sigma(x_3)^2 + NA_3^2]b_{13.24} + [\Sigma(x_3x_4) + NA_3A_4]b_{14.23}$$

$$\text{IV}\quad \Sigma(x_1x_4) + NA_1A_4 = NA_4a + [\Sigma(x_2x_4) + NA_2A_4]b_{12.34} + [\Sigma(x_3x_4) + NA_3A_4]b_{13.24} + [\Sigma(x_4)^2 + NA_4^2]b_{14.23}.$$

If we now divide through by N, and substitute p_{12} for $\frac{\Sigma x_1x_2}{N}$, σ_2^2 for $\frac{\Sigma(x_2^2)}{N}$, and similar symbols for other mean products and mean squares, the normal equations become

$$\text{I} \quad A_1 = a + A_2 b_{12.34} + A_3 b_{13.24} + A_4 b_{14.23}$$

$$\text{II} \quad p_{12} + A_1 A_2 = A_2 a + (\sigma_2^2 + A_2^2) b_{12.34} + (p_{23} + A_2 A_3) b_{13.24} + (p_{24} + A_2 A_4) b_{14.23}$$

$$\text{III} \quad p_{13} + A_1 A_3 = A_3 a + (p_{23} + A_2 A_3) b_{12.34} + (\sigma_3^2 + A_3^2) b_{13.24} + (p_{34} + A_3 A_4) b_{14.23}$$

$$\text{IV} \quad p_{14} + A_1 A_4 = A_4 a + (p_{24} + A_2 A_4) b_{12.34} + (p_{34} + A_3 A_4) b_{13.24} + (\sigma_4^2 + A_4^2) b_{14.23}.$$

These four simultaneous equations may now be reduced to three. We multiply equation I, throughout, by A_2, and subtract the result from equation II; we then multiply equation I by A_3, and subtract the result from equation III; we then multiply equation I by A_4, and subtract the result from equation IV. All the terms containing A's are thus eliminated and we obtain the three normal equations

$$p_{12} = \sigma_2^2 b_{12.34} + p_{23} b_{13.24} + p_{24} b_{14.23}$$
$$p_{13} = p_{23} b_{12.34} + \sigma_3^2 b_{13.24} + p_{34} b_{14.23}$$
$$p_{14} = p_{24} b_{12.34} + p_{34} b_{13.24} + \sigma_4^2 b_{14.23}.$$

Inserting the observed values of the p's and the σ's, these are solved for the coefficients b. The value a may then be obtained by inserting the values of the A's and the b's in the equation

$$A_1 = a + A_2 b_{12.34} + A_3 b_{13.24} + A_4 b_{14.23}.$$

SOLUTION OF THE NORMAL EQUATIONS

The task of solving the normal equations is not a difficult one in most of the cases presented to the economic statistician. If there are only two or three unknowns the corresponding number of normal equations may be solved by simple algebraic methods. Even with three equations, however, it is advisable to employ a systematic procedure, and with more than three equations this is imperative. Such systematic methods of solving the simultaneous equations which are met with in connection with the method of least squares have been worked out by Gauss and by

Doolittle. The latter method, which is perhaps the more convenient for general usage, is demonstrated below.

The coefficients of the unknowns in the normal equations are always symmetrical with respect to the principal diagonal. Thus in securing the most probable values of the constants in the equation

$$Y = aW_1 + bW_2 + cW_3 + dW_4,$$

we have the four normal equations

$$\begin{aligned}
a\Sigma(W_1^2) + b\Sigma(W_1W_2) + c\Sigma(W_1W_3) + d\Sigma(W_1W_4) - \Sigma(W_1Y) &= 0\\
a\Sigma(W_1W_2) + b\Sigma(W_2^2) + c\Sigma(W_2W_3) + d\Sigma(W_2W_4) - \Sigma(W_2Y) &= 0\\
a\Sigma(W_1W_3) + b\Sigma(W_2W_3) + c\Sigma(W_3^2) + d\Sigma(W_3W_4) - \Sigma(W_3Y) &= 0\\
a\Sigma(W_1W_4) + b\Sigma(W_2W_4) + c\Sigma(W_3W_4) + d\Sigma(W_4^2) - \Sigma(W_4Y) &= 0
\end{aligned}$$

The symmetrical arrangement about the diagonal, when Y-terms are neglected, is obvious. Starting with any term on the principal diagonal, we have the same coefficients directly above as to the left. Thus, above the diagonal term in which the coefficient $\Sigma(W_3^2)$ appears, we have the coefficients $\Sigma(W_2W_3)$ and $\Sigma(W_1W_3)$. The same coefficients are found to the left of the given diagonal term, and on the same line. For the purposes of solution, therefore, the terms to the left of each diagonal entry may be omitted, and we may put the remaining terms of the normal equations in the form

$$\begin{aligned}
a\Sigma(W_1^2) + b\Sigma(W_1W_2) + c\Sigma(W_1W_3) + d\Sigma(W_1W_4) - \Sigma(W_1Y)\\
+ b\Sigma(W_2^2) + c\Sigma(W_2W_3) + d\Sigma(W_2W_4) - \Sigma(W_2Y)\\
+ c\Sigma(W_3^2) + d\Sigma(W_3W_4) - \Sigma(W_3Y)\\
+ d\Sigma(W_4^2) - \Sigma(W_4Y).
\end{aligned}$$

THE DOOLITTLE METHOD

The Doolittle method may be illustrated with reference to the following normal equations:

$$\begin{aligned}
8.3564a + 2.790b + 2.932c + 47.967 &= 0\\
2.790a + 6.6645b + 2.063c + 62.039 &= 0\\
2.932a + 2.063b + 7.7893c + 47.519 &= 0.
\end{aligned}$$

Putting these, for the purposes of the solution, in the abbreviated form given above, we have

$$\begin{aligned} 8.3564a + 2.790b &+ 2.932c + 47.967 \\ + 6.6645b &+ 2.063c + 62.039 \\ &+ 7.7893c + 47.519. \end{aligned}$$

We wish to solve these for the constants a, b, and c. All the work of computation, with the necessary checks, is shown in the following table:

TABLE B

Solution of Normal Equations by the Doolittle Method

Line	(1) Reciprocals	(2) a	(3) b	(4) c	(5)	(6) s
I		8.3564	2 790	2 932	47 967	62 0454
II			6 6645	2 063	62 039	73 5565
III				7 7893	47 519	60 3033
1		8 35640	2 790	2 932	47 967	62 0454
2	− .11966876	− 1 00000	− 333876	− 350869	− 5 740151	− 7.424896 check
3			6 6645	2 063	62 039	73 5565
4			− 931514	− 978924	− 16 015030	− 20 715470
5			5 732986	1 084076	46 023970	52 841030 check
6	− .17442917		− 1 000000	− 189094	− 8 027923	− 9 217017 check
7				7 7893	47 519	60 3033
8				− 1 028748	− 16 830133	− 21 769807
9				− 204992	− 8 702857	− 9 991922
10				6 555560	21 986010	28 541571 check
11	− .15254227			− 1 000000	− 3 353796	− 4 353796 check

Back Solution

c	b	a
− 3 353796	− 8 027923	− 5 740151
− 3.353796	+ 634183	+ 2 468592
	− 7.393740	+ 1 176743
		− 2.094816

$$a = -2.094816$$
$$b = -7.393740$$
$$c = -3.353796$$

Check:

Equation I:

$$8.3564a + 2.790b + 2.932c = -47.967.$$

Substituting the given values,

$$8\ 3564(-2.094816) + 2.790(-7.393740) + 2.932(-3.353796) = -47.966985.$$

Explanation.— The coefficients of the unknown quantities, a, b, and c, are listed in the designated columns. The known term in each normal equation is listed in column (5). (The sign of this known term, it should be noted, is that which it would have when the entire expression, of which it is one term, is equated to zero.) Column s is employed as a check. The value in column s, in each of the lines I, II, and III, is the algebraic sum of the known values in the given normal equation. In securing this sum the coefficients to the left of the diagonal, which have been omitted from the table as it stands, must be included.

The following is a summary of the procedure in solving the normal equations:

1. In line (1) write normal equation I.

2. In line (2), column (1), write the reciprocal of the value in line (1), column (2), *with sign changed.* (This is the reciprocal of the coefficient of a.) Multiply each item in line (1) by this reciprocal, entering the products in the corresponding columns in line (2). [The algebraic sum of the items in columns (2), (3), (4), and (5) of line (2) should equal the value in column (6).] This operation has eliminated the unknown a, by expressing it in terms of b and c. [The -1 in line (2), column (2), has been included only to facilitate the checking process. The same is true in lines (6) and (11).] A heavy line may be drawn across the table below line (2).

3. Write normal equation II in line (3).

4. Multiply by the coefficient of b in line (2) (i.e., $-.333876$) the items in columns (3), (4), (5), and (6) in line (1). Enter the products in the corresponding columns of line (4).

5. Add lines (3) and (4), entering the sums in line (5). [The algebraic sum of the items in columns (3), (4), and (5) of line (5) should equal the value in column (6).]

6. In column (1), line (6), enter the reciprocal of the value in column (3), line (5), *reversing the sign.* Multiply each term in line (5) by this reciprocal, entering the products in line (6). [The sum of the items in columns (3), (4), and (5) of line (6) should equal the value in column (6).] This operation has eliminated the unknown b, by expressing it in terms of c. A heavy line may be drawn across the table below line (6).

7. Write normal equation III in line (7).

8. Multiply by the coefficient of c in line (2) (i.e., $-.350869$)

the items in columns (4), (5), and (6) of line (1). Enter the products in the corresponding columns of line (8).

9. Multiply by the coefficient of c in line (5) (i.e., $-.189094$) the items in columns (4), (5), and (6) of line (5). Enter the products in the corresponding columns of line (9).

10. Add lines (7), (8), and (9), entering the sums in line (10). [The algebraic sum of the items in columns (4) and (5) of line (10) should equal the value in column (6).]

11. In column (1), line (11), enter the reciprocal of the value in column (4) of line (10), *reversing the sign.* Multiply each term in line (10) by this reciprocal, entering the products in line (11). [The algebraic sum of the items in columns (4) and (5) of line (11) should equal the value in column (6).] This operation gives the value of c, which is found in column (5) of line (11). A heavy line may be drawn across the table below line (11).

Were there additional unknowns, as d and e, this last operation would have given c as a function of d and e and it would be necessary to carry the process still further, repeating the steps taken above. The next operation would be to bring down the fourth normal equation, entering it in line (12). Then the coefficients of d in lines (2), (6), and (11) would be used to multiply the necessary items in lines (1), (5), and (10), the products being entered in lines (13), (14), and (15). The sum of the items in lines (12), (13), (14), and (15) would be entered in line (16) and checked by the item in the s column. Multiplying through by the reciprocal of the coefficient of d in line (16), with sign reversed, the value of d would be obtained in terms of e. The value of e would be derived in a similar fashion.

The checks on these various operations have been indicated in the table. The testing of the results at each step reduces the possibility of error to a minimum.

The back solution presents no difficulties. We have, from line (11),

$$c = -3.353796,$$

from line (6)

$$b = -.189094c - 8.027923,$$

from line (2)

$$a = - .333876b - .350869c - 5.740151.$$

[The items in column (6) are inserted merely as checks. The items – 1.000000 which appear in lines (2), (6), and (11) are inserted to assist in the checking.]

The computations involved in the back solution appear in the table.

A final check is afforded by inserting the values secured by this process in one of the normal equations. This check, as carried out for equation I, is shown below the table.

REFERENCES

Brunt, David, *The Combination of Observations.*

Huntington, E. V., "Curve Fitting by the Method of Least Squares and the Method of Moments" (in Rietz, H. L., ed., *Handbook of Mathematical Statistics,* 62–70).

Merriman, Mansfield, *The Method of Least Squares.*

Tolley, H. R. and Ezekiel, M. J. B., "A Method of Handling Multiple Correlation Problems," *Journal of the American Statistical Association.* Dec., 1923.

Weld, L. D., *Theory of Errors and Least Squares.*

Whittaker, E. T. and Robinson, G., *The Calculus of Observations* (209–259).

Wright, T. W. and Hayford, J. F., *Adjustment of Observations.*

APPENDIX B

DERIVATION OF FORMULAS FOR MEAN AND STANDARD DEVIATION OF THE BINOMIAL DISTRIBUTION [1]

For convenience we put the binomial in the form $(q + p)^n$, where q = probability of a failure, p = probability of a success, and $q + p = 1$. Expanding the binomial, we have

$$(q + p)^n = q^n + nq^{n-1}p^1 + \frac{n(n-1)}{1 \cdot 2} q^{n-2}p^2 + \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} q^{n-3}p^3 + \ldots + p^n.$$

The terms of this expansion indicate, in order, the probable frequencies of no successes, 1 success, 2 successes, 3 successes, and so on, to n successes. A frequency table of the familiar type may be constructed from these materials.

The items in col. (2) of Table C constitute the terms of the binomial expansion. Their sum is thus equal to $(q + p)^n$, which is, by definition, equal to 1. The items in col. (3), added in order, give

$$nq^{(n-1)}p^1 + n(n-1)q^{n-2}p^2 + \frac{n(n-1)(n-2)}{1 \cdot 2} q^{n-3}p^3 + \frac{n(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3} q^{n-4}p^4 + \ldots + np^n.$$

Since the factors n and p appear in each of these terms, this reduces to

$$np\left[q^{n-1} + (n-1)(q^{n-2}p^1) + \frac{(n-1)(n-2)}{1 \cdot 2} q^{n-3}p^2 + \frac{(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3} q^{n-4}p^3 + \ldots p^{n-1}\right].$$

[1] These derivations are adapted from the proof given by D. C. Jones in *A First Course in Statistics*. London, Bell & Sons, 1921, 143–145.

TABLE C

Derivation of Mean and Standard Deviation of the Binomial Distribution

(1) *Number of successes* m	(2) *Frequency* f	(3) fm	(4) fm^2
0	q^n	0	0
1	$nq^{n-1}p^1$	$nq^{n-1}p^1$	$nq^{n-1}p^1$
2	$\frac{n(n-1)}{1\cdot 2}q^{n-2}p^2$	$n(n-1)q^{n-2}p^2$	$2n(n-1)q^{n-2}p^2$
3	$\frac{n(n-1)(n-2)}{1\cdot 2\cdot 3}q^{n-3}p^3$	$\frac{n(n-1)(n-2)}{1\cdot 2}q^{n-3}p^3$	$\frac{3n(n-1)(n-2)}{1\cdot 2}q^{n-3}p^3$
4	$\frac{n(n-1)(n-2)(n-3)}{1\cdot 2\cdot 3\cdot 4}q^{n-4}p^4$	$\frac{n(n-1)(n-2)(n-3)}{1\cdot 2\cdot 3}q^{n-4}p^4$	$\frac{4n(n-1)(n-2)(n-3)}{1\cdot 2\cdot 3}q^{n-4}p^4$
.....	.....	.....	.....
.....	.....	.....	.....
.....	.....	.....	.....
.....	.....	.....	.....
.....	.....	.....	.....
.....	.....	.....	.....
n	p^n	np^n	n^2p^n
Total	1	np	$np[1+p(n-1)]$

But the terms within brackets, following np, represent the expansion of the binomial $(q + p)^{n-1}$. Since $q + p = 1$, the sum of these terms is 1. Accordingly the sum of the items in col. (3) reduces to

$$np(q + p)^{n-1} = np.$$

For the mean of this distribution we have

$$M = \frac{\Sigma(fm)}{\Sigma(f)} = \frac{np}{1} = np.$$

Adding the items in col. (4) in order, we have

$$nq^{n-1}p^1 + 2n(n-1)q^{n-2}p^2 + \frac{3n(n-1)(n-2)}{1 \cdot 2} q^{n-3}p^3$$
$$+ \frac{4n(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3} q^{n-4}p^4 + \ldots + n^2p^n$$
$$= np\Big[q^{n-1} + 2(n-1)q^{n-2}p^1 + \frac{3(n-1)(n-2)}{1 \cdot 2} q^{n-3}p^2$$
$$+ \frac{4(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3} q^{n-4}p^3 + \ldots + np^{n-1}\Big].$$

The terms within brackets may be broken into two groups, giving

$$np\Big[\Big\{q^{n-1} + (n-1)q^{n-2}p^1 + \frac{(n-1)(n-2)}{1 \cdot 2} q^{n-3}p^2$$
$$+ \frac{(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3} q^{n-4}p^3 + \ldots + p^{n-1}\Big\}$$
$$+ \Big\{(n-1)q^{n-2}p^1 + \frac{2(n-1)(n-2)}{1 \cdot 2} q^{n-3}p^2$$
$$+ \frac{3(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3} q^{n-4}p^3 + \ldots + (n-1)p^{n-1}\Big\}\Big].$$

The terms within the first of these two groups constitute the expansion of the binomial $(q + p)^{n-1}$. These terms may be replaced by that binomial; the second group of terms may be simplified, since they contain the common factors $n - 1$ and p. These operations give us

$$np\Big[(q+p)^{n-1}+(n-1)p\Big\{q^{n-2}+(n-2)q^{n-3}p^1 + \frac{(n-2)(n-3)}{1\cdot 2}q^{n-4}p^2+\ldots+p^{n-2}\Big\}\Big].$$

The second group of terms, thus simplified, is seen to be $(n-1)p$ multiplied by the expansion of the binomial $(q+p)^{n-2}$. Thus we have, as the sum of the items in col. (4) of the preceding table,

$$np[(q+p)^{n-1}+(n-1)p(q+p)^{n-2}].$$

But since $q+p=1$, $(q+p)^{n-1}=1$ and $(q+p)^{n-2}=1$. Accordingly, the total of col. (4) becomes

$$np[1+p(n-1)].$$

As a general formula for the standard deviation, in squared form, we have

$$\sigma^2=\frac{\Sigma fm^2}{N}-c^2$$

where c is the difference between the mean of the distribution and the arbitrary origin. In the present instance, the origin is at 0, or "no successes," and c is equal to the mean, or np. N is equal to $\Sigma(f)$, or 1, in this case. Thus the standard deviation of the binomial distributions given by

$$\begin{aligned}\sigma^2 &= np[1+p(n-1)]-n^2p^2\\ &= np[np+(1-p)]-n^2p^2\\ &= n^2p^2+np(1-p)-n^2p^2\\ &= np(1-p)\\ &= npq\\ \sigma &= \sqrt{npq}.\end{aligned}$$

APPENDIX C

DERIVATION OF THE STANDARD ERROR OF THE ARITHMETIC MEAN

We have made n random, hence independent, observations on a given variable. The respective observations may be represented by X_1, X_2, X_3 . . . X_n. Representing the sum of the n observations by W, we have

$$W = X_1 + X_2 + X_3 + \ldots + X_n. \tag{1}$$

Additional samples are now taken until we have N values of X_1, N values of X_2, etc., and hence N values of the sum, W. We have N samples, therefore, of n observations each. The mean values, which we may represent by barred letters, stand in the same relationship of equality:

$$\overline{W} = \overline{X}_1 + \overline{X}_2 + \overline{X}_3 + \ldots + \overline{X}_n. \tag{2}$$

Using small letters (w, x_1, x_2, etc.) to define deviations of the actual observations from these mean values, we may write, for any given sample, or series of observations,

$$w = x_1 + x_2 + x_3 + \ldots + x_n. \tag{3}$$

Squaring the two sides of this equation, we have

$$\begin{aligned} w^2 = {} & x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2 + 2x_1x_2 + 2x_1x_3 + \ldots \\ & + 2x_1x_n + 2x_2x_3 + \ldots + 2x_2x_n + \ldots \\ & + 2x_3x_n + \ldots . \end{aligned} \tag{4}$$

Each term on the right-hand side of (3) will appear in squared form in (4), and there will also appear product terms of the form $2x_1x_2$ corresponding to all possible pairings of the terms on the right-hand side.

The next step involves the summation of the equations of type (4), derived from the N samples, and division

throughout by N. Each product term, when thus summed and divided by N, will be of the form

$$\frac{2\Sigma x_1 x_2}{N}$$

This, with the modification introduced by the factor 2, resembles the familiar mean product, $\frac{\Sigma xy}{N}$, encountered in correlation procedure. This mean product, we have seen, has a value of zero when the variables x and y are uncorrelated. But, by hypothesis, the observations that have given us x_1, x_2, x_3, etc., are independent of one another, and hence these variables are uncorrelated. Accordingly, each of the product terms, derived when N equations corresponding to (4) above are summed and divided by N, is equal to zero. The process of summation and division gives us, therefore,

$$\frac{\Sigma w^2}{N} = \frac{\Sigma x_1{}^2}{N} + \frac{\Sigma x_2{}^2}{N} + \frac{\Sigma x_3{}^2}{N} + \ldots + \frac{\Sigma x_n{}^2}{N} \tag{5}$$

or

$$\sigma_w{}^2 = \sigma_1{}^2 + \sigma_2{}^2 + \sigma_3{}^2 + \ldots + \sigma_n{}^2. \tag{6}$$

If all the observations relate to the same universe (i.e., if the samples are all drawn from the same parent population), which is true, by hypothesis, the standard deviations appearing in the right-hand member of equation (6) are equal to one another and to the standard deviation of the population. Accordingly, using σ to represent that standard deviation, we have

$$\sigma_w{}^2 = n\sigma. \tag{7}$$

The next argument, that leads directly to the desired measurement, follows precisely these steps, which have been given in the above form to indicate the reasoning involved. It starts, however, with a variant form of equation (3). Dividing that equation throughout by n, we have

$$\frac{w}{n} = \frac{x_1}{n} + \frac{x_2}{n} + \frac{x_3}{n} + \ldots + \frac{x_n}{n}. \tag{8}$$

Working with the variables $\frac{w}{n}$, $\frac{x_1}{n}$, $\frac{x_2}{n}$, etc., just as we have done with w, x_1, x_2, etc., we may go through the operations represented by equations (4), (5), and (6), above. The product terms disappear, as in passing from (4) to (5). In the process of squaring the term $\frac{w}{n}$ is treated as an entity; the sum of the squared values is thus $\Sigma\left(\frac{w}{n}\right)^2$. Numerator and denominator of each of the terms of type $\frac{x_1}{n}$ are squared separately, however, and the sum is of the form $\frac{\Sigma x^2}{n^2}$. Division throughout by N then gives the quantities appearing in equation (9), which corresponds to equation (6).

$$\sigma_{\frac{w}{n}}^2 = \frac{\sigma_1^2}{n^2} + \frac{\sigma_2^2}{n^2} + \frac{\sigma_3^2}{n^2} + \ldots + \frac{\sigma_n^2}{n^2}. \tag{9}$$

Since all observations relate to the same universe, this reduces to

$$\sigma_{\frac{w}{n}}^2 = \frac{n\sigma^2}{n^2}. \tag{10}$$

From this

$$\sigma_{\frac{w}{n}} = \frac{\sigma}{\sqrt{n}}. \tag{11}$$

But w is the sum of n observations drawn from a universe having a standard deviation of σ, and $\frac{w}{n}$ is the mean of these observations. $\sigma_{\frac{w}{n}}$ is the standard deviation of a distribution of arithmetic means, corresponding to the familiar symbol σ_M. This is the desired expression for the standard error of the arithmetic mean, appropriate for use when the σ of the population is known. Where σ is estimated from the standard deviation of a sample, accuracy is increased by using $\sqrt{n-1}$ rather than $\sqrt{n}$ in the denominator of the right-hand member of (11).

APPENDIX D

ILLUSTRATING THE MEASUREMENT OF TREND BY A MODIFIED EXPONENTIAL CURVE, A GOMPERTZ CURVE AND A LOGISTIC CURVE

The discussion in Chapter VII of mathematical functions suitable for use in measuring the secular trends of time series dealt with types required in ordinary practice. We here discuss briefly three other types suited to the measurement of long-term movements in economic and business series.

The Modified Exponential Curve

An exponential curve, which plots as a straight line on ratio paper, is a suitable measure of trend for a series that is increasing or decreasing at a constant rate, that is, one that shows constancy of relative growth. The figures defining the successive trend values of a series of this type constitute a geometric progression. The trends of certain economic series that depart from constancy of relative growth may be accurately defined by a simple modification of the exponential curve. This is the case when the observed values may be transformed, by the addition (or subtraction) of a constant magnitude, to a series closely approximating such a geometric progression.

If we represent by K the constant magnitude that is to be added (algebraically) to each observed value in effecting the desired transformation, the task of fitting the trend line involves the following steps:

Determination of K.
Correction of observed values by K, to obtain the modified series.
Fitting an exponential curve to the modified series, and computation of trend values of the modified series.

Correction of trend values of the modified series by K to obtain trend values of original series.

If y represents the ordinates of trend of the original series and x represents time, the equation to the desired line of trend may be put in the form

$$y = ab^x - K$$

where K is the correction factor noted above and a and b are constants to be determined by fitting an exponential curve to the modified series. The procedure may be illustrated with reference to the data in Table D.

TABLE D

Illustrating the Fitting of a Modified Exponential Curve Production of Rayon Filament Yarn in the United States, 1920–1931 [1]

(Data in thousands of pounds)

(1) Year	(2) Original series (observed)	(3) Group mean	(4) Modified series (2) + K	(5) Trend values, modified series	(6) Trend values, original series (5) − K
1920	10,125		27,669	29,108	11,564
1921	14,986	M_1 =	32,530	34,363	16,819
1922	24,067	21,034.25	41,611	40,565	23,021
1923	34,959		52,503	47,888	30,344
1924	36,328		53,872	56,532	38,988
1925	51,049	M_2 =	68,593	66,736	49,192
1926	62,693	56,406.25	80,237	78,782	61,238
1927	75,555		93,099	93,003	75,459
1928	97,232		114,776	109,790	92,246
1929	121,399	M_3 =	138,943	129,608	112,064
1930	127,333	124,210.75	144,877	153,003	135,459
1931	150,879		168,423	180,621	163,077

In employing this method we approximate K empirically by breaking the observed series into three parts, representing equal periods of time, and determining the mean of the

[1] Data from Textile Economics Bureau.

observations for each period. We may designate these means, in chronological order, by M_1, M_2, and M_3. The desired value, K, is given by

$$K = [M_2^2 - (M_1 \times M_3)] \div [(M_1 + M_3) - 2M_2].$$

If the observed series constitute a geometric progression the value of K will be zero; if the *addition* of a constant magnitude to the members of the original series will yield a series approximating a geometric progression, K will be positive; if the *subtraction* of a constant amount from the observed values will yield a series approximating a geometric progression, K will be negative. (In practice, K is given the sign obtained by the employment of the method described above, and then added algebraically to the observed series.)

In the present case we have

$$K = [(56{,}406.25)^2 - (21{,}034.25 \times 124{,}210.75)] \div [(21{,}034.25 + 124{,}210.75) - (2 \times 56{,}406.25)]$$
$$= +\,17{,}544.$$

Adding this amount to each of the values recorded in col. (2) of Table D, we obtain the modified series in col. (4). In fitting an exponential curve to the modified series, it is desirable to use logarithms, that is, to solve the constants in an equation of the type $\log y = \log a + (\log b)x$. This procedure was explained in Chapter VII. For $\log a$ of this curve we obtain 4.824359, and for $\log b$ .072068. (The origin is at 1925.) The antilogarithms of the series of trend values thus obtained are given in col. (5). These define the trend of the modified series. Subtracting K (algebraically) from these values we obtain the trend values of the original series, which appear in col. (6).

The original series measuring production of rayon filament yarn and the modified exponential curve fitted to this series are shown graphically in Fig. A.

It is essential that the three M's used in the determina-

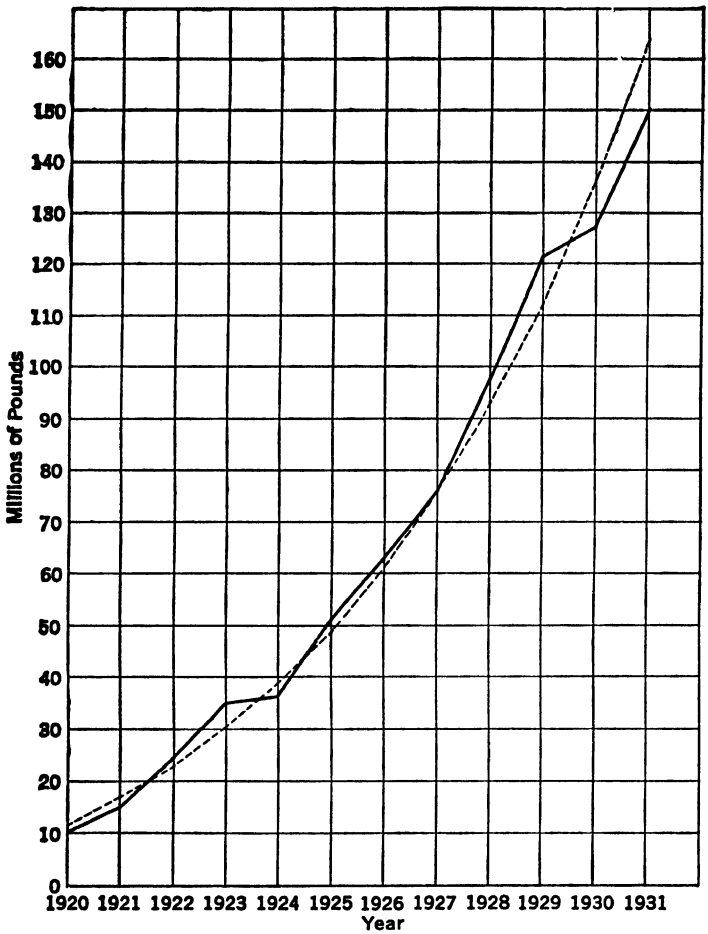


FIG. A. — Total Production of Rayon Filament Yarn in the United States, 1920–1931, with Modified Exponential Trend

tion of K relate to equal numbers of observations and that the midpoints, in time, of the three periods be equidistant. In the above example the number of years included in the period is a multiple of three, and no difficulty arises. If the number of years included is not a multiple of three, intervals that overlap slightly may be employed. For example, if our series had run from 1920 to 1932, the three averages might

have been derived, respectively, from the five-year periods 1920–1924, 1924–1928, 1928–1932. These would center, respectively, at 1922, 1926, and 1930, and would thus be equidistant in time from one another. Alternatively, if monthly data are available, division of the total period into three equal parts may be facilitated by using a time-unit of 4 or 8 months, rather than 12 months.

The Gompertz Curve

The Gompertz curve, which has important uses in actuarial science, has had some application in the study of economic and business trends.[1] The term "growth curve" is applicable to it, since it portrays a process of cumulative expansion to a maximum value. This expansion proceeds by decreasing absolute increments in the later stages, but continues to the end without retrogression. It may not be assumed that this form of growth is typical of all industrial development, but the curve has value as an empirical representation of certain trend movements.

For the purpose of fitting, the equation to the curve is transformed from the natural form

$$y = ab^{c^x}$$

to the logarithmic form

$$\log y = \log a + (\log b)c^x.$$

When fitted to an appropriate set of observations, measuring the expansion of an industry or the growth of an economic element, $\log a$ is the logarithm of the maximum value — the *ceiling* that the curve approaches. The second term measures the amount by which the trend value at a given time falls short of this maximum, an amount that diminishes, of course, with the passage of time. (The series for which this curve is an appropriate measure of trend will be expanding by decreasing amounts in the later stages of the period covered, and c, derived in the manner indicated below, will have a value between zero and unity.) The origin on

[1] Cf. Raymond B. Prescott, "Law of Growth in Forecasting Demand," *Journal of the American Statistical Association*, Dec., 1922.

the x-scale (time) is taken at the year to which the first entry relates.

The method employed in fitting this curve is an approximative one, since the least squares procedure in customary form is not applicable. Here, as in the preceding example, the series is broken into three equal portions. The sum of the logarithms of the observations in each of these segments is obtained; from these sums, and the differences between them, the necessary constants may be computed. The method is illustrated with reference to the data of rayon production for the years 1920–1937, which appear in Table E.

TABLE E

Computation of Quantities Required in the Fitting of a Gompertz Curve

Production of Rayon Filament Yarn in the United States, 1920–1937

(Data in thousands of pounds)

Year	*Rayon production* y	*Log y*	*Sub-totals*	*First differences*
1920	10,125	4.0053950		
1921	14,986	4.1756857		
1922	24,067	4.3814220		
1923	34,959	4.5435590	$S_1 =$	
1924	36,328	4.5602415	26.374290	
1925	51,049	4.7079872		
1926	62,693	4.7972191		$d_1 = S_2 - S_1$
1927	75,555	4.8782632		$= 3.656786$
1928	97,232	4.9878092		
1929	121,399	5.0842151	$S_2 =$	
1930	127,333	5.1049409	30.031076	
1931	150,879	5.1786288		
1932	134,670	5.1292709		$d_2 = S_3 - S_2$
1933	213,498	5.3293938		$= 2.095138$
1934	208,321	5.3187331		
1935	257,557	5.4108734	$S_3 =$	
1936	277,626	5.4434601	32.126214	
1937	312,236	5.4944829		
		88.5315830		

We may use n to define the number of terms entering into each of the three sub-totals (in the present example $n = 6$); the sub-totals are represented, in chronological order, by S_1, S_2, and S_3; the first differences between the sub-totals are represented by d_1 and d_2. [1] We use these quantities in solving for the three constants c, $\log b$ and $\log a$. The general relations from which these values are determined are the following:

$$c^n = \frac{d_2}{d_1}$$

$$\log b = \frac{d_1(c-1)}{(c^n-1)^2}$$

$$\log a = \frac{1}{n}\left\{ S_1 - \frac{d_1}{c^n - 1} \right\}.$$

Inserting the proper quantities, we have

$$c^n = c^6 = \frac{2.095138}{3.656786} = .572945$$

$$c = \sqrt[6]{.572945} = .911351$$

$$\log b = \frac{3.656786 \times -.088649}{(.572945-1)^2} = -1.777493$$

$$\log a = \frac{1}{6}\left\{ 26.374290 - \frac{3.656786}{.572945 - 1} \right\}$$

$$= 5.822848.$$

The required equation is, therefore,

$$\log y = 5.822848 - 1.777493(.911351^x)$$

in which x relates to deviations from an origin at the position of the first term.

Substituting in this trend equation the values of x given in Table F, logarithms of the trend values are obtained. The corresponding natural numbers define the course of the line of trend. The method of calculation is indicated in Table F.

[1] The condition, previously noted, that the series to which the curve is to be fitted be one that is expanding by decreasing increments in the later stages of the period covered, is met when d_2 is less than d_1.

TABLE F

Illustrating the Computation of Ordinates of Trend of a Gompertz Curve Fitted to Data of Rayon Production, 1920–1937

(1)	(2)	(3)	(4)	(5)	(6)
Year	x	c^x	$(\log b)c^x$	$\log y$ (4) + $\log a$	y *Anti-log of* (5) *(in thousands of pounds)*
1920	0	1.000000	− 1.777493	4.045355	11,101
1921	1	0.911351	− 1.619920	4.202928	15,956
1922	2	0.830560	− 1.476315	4.346533	22,209
1923	3	0.756932	− 1.345441	4.477407	30,020
1924	4	0.689830	− 1.226168	4.596680	39,508
1925	5	0.628677	− 1.117469	4.705379	50,743
1926	6	0.572945	− 1.018408	4.804440	63,744
1927	7	0.522154	− 0.928125	4.894723	78,474
1928	8	0.475865	− 0.845847	4.977001	94,842
1929	9	0.433681	− 0.770865	5.051983	112,715
1930	10	0.395235	− 0.702527	5.120321	131,923
1931	11	0.360198	− 0.640249	5.182599	152,265
1932	12	0.328267	− 0.583492	5.239356	173,523
1933	13	0.299166	− 0.531765	5.291083	195,471
1934	14	0.272645	− 0.484625	5.338223	217,883
1935	15	0.248475	− 0.441663	5.381185	240,539
1936	16	0.226448	− 0.402510	5.420338	263,231
1937	17	0.206374	− 0.366830	5.456018	285,771
1947	27	0.081566	− 0.144983	5.677865	476,283
1957	37	0.032238	− 0.057303	5.765545	582,834
1967	47	0.012741	− 0.022647	5.800201	631,245

The original data and the Gompertz curve fitted to them are shown graphically in Fig. B.

The ceiling to this curve is set by the constant a, which has a value of approximately 665,000,000 pounds. This indicates that if the extrapolation of the trend of rayon production from 1920 to 1937, as measured by a Gompertz curve, accurately defines the future course of production, the maximum output to be expected is 665 million pounds per year. (It need hardly be pointed out that this extrapolation involves some doubtful assumptions, and that no mystic significance is to be attached to it.) The years to

which the present data relate were years of rapid expansion in the industry. The slackening of the rate of increase, which is to be expected in a mature industry, had not become marked by 1937. In order that the nature of the curve may be clear, extrapolated values for 1947, 1957, and 1967 are given in the table, and the projection of the trend is

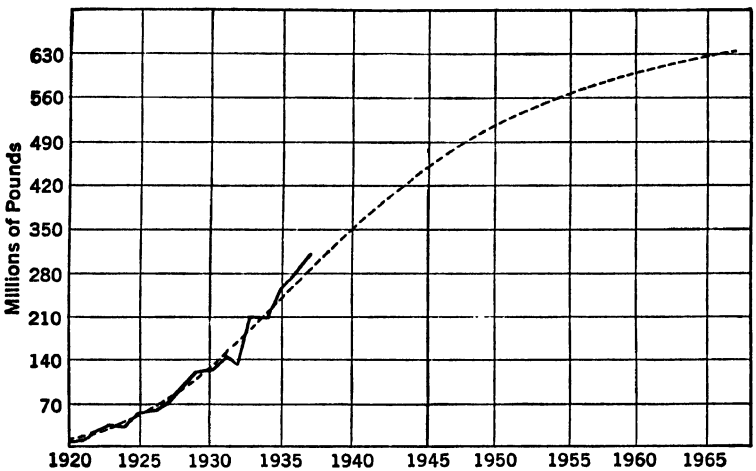


FIG. B. — Total Production of Rayon Filament Yarn in the United States, 1920–1937, with Gompertz Trend Line Extrapolated to 1967

shown in Fig. B. After 1947, and still more conspicuously after 1957, the curve shows a notable dampening in the rate of expansion. We may not say that the industry will actually follow this course. In particular, the asymptote *a* may be expected to change, as conditions affecting the industry and the demand for its products vary in the future. Within the limits of the observations, however, the Gompertz curve serves as a satisfactory measure of trend.

THE LOGISTIC CURVE

The logistic curve, sometimes termed the Pearl-Reed growth curve because of the extensive use made of it in population studies by Raymond Pearl and L. J. Reed, resembles somewhat the Gompertz curve discussed above.

TABLE G

Computation of Quantities Required in the Fitting of a Logistic Curve to Data of Railroad Mileage Operated in the United States, by Five-Year Intervals, 1850–1935

(1)	(2)	(3)	(4)	(5)
Year	*Miles of railroad operated* y	$\frac{100{,}000{,}000}{y}$	*Sub-totals*	*First differences*
1850	9,021	11,085		
1855	18,374	5,442		
1860	30,626	3,265		
1865	35,085	2,850	$S_1 = 25{,}882$	
1870	52,922	1,890		
1875	74,096	1,350		
1880	93,262	1,072		$d_1 = S_2 - S_1$
1885	128,320	779		$= -21{,}849$
1890	156,404	639	$S_2 = 4{,}033$	
1895	177,746	563		
1900	192,556	519		
1905	216,974	461		
1910	240,831	415		$d_2 = S_3 - S_2$
1915	257,569	388		$= -1{,}679$
1920	259,941	385	$S_3 = 2{,}354$	
1925	258,631	387		
1930	260,440	384		
1935	252,930	395		

It represents a modified geometric progression, the growth of a series that tends to decrease as it approaches some specified limit. Like the Gompertz curve it may be used as an empirical approximation to the trends of certain economic series. Extrapolations are subject, of course, to the same uncertainties that attach to projections of other empirically derived trend lines.

A form of this curve adapted to use as a measure of trend is defined by the equation

$$\frac{1}{y} = a + bc^x.$$

This, it will be noted, is the equation to a modified exponential curve, except that the dependent variable is $\frac{1}{y}$, rather than y. (The symbols here used for the constants differ somewhat from those employed in treating the modified exponential curve.) A method of fitting somewhat similar to those employed in the preceding examples may be employed, with necessary modifications required by the use of reciprocals of y. The method may be discussed with reference to the data of railroad mileage in Table G. Computations are facilitated by multiplying the reciprocals of y by a suitable power of 10, as is done in col. (3) of this table.

As in the two preceding illustrations, the observations are divided, chronologically, into three equal groups. Group sub-totals and the first differences between these sub-totals are computed. The symbol n is used for the number of terms in each of these sub-groups. The origin of the x-scale (time) is set at the date of the first observation. The time unit here employed is five years.

The constants in the desired equation may be derived from the following relations.

$$c^n = \frac{d_2}{d_1}$$

$$b = \frac{d_1(c-1)}{(c^n-1)^2}$$

$$a = \frac{1}{n}\left\{ S_1 - \frac{d_1}{c^n - 1} \right\}.$$

Substituting the given values, we have

$$c^n = c^6 = \frac{-1{,}679}{-21{,}849} = +.076846$$

$$c = \sqrt[6]{.076846} = .652034$$

$$b = \frac{-21{,}849(-.347966)}{(.076846-1)^2} = +8{,}921.14$$

$$a = \frac{1}{6}\left\{ 25{,}882 - \frac{-\ 21{,}849}{(.076846 - 1)} \right\} = +\ 369.04.$$

These results relate to initial observations which have been modified by the multiplication of $\frac{1}{y}$ by 100,000,000. The desired equation is, therefore,

$$\frac{100{,}000{,}000}{y} = 369.04 + 8{,}921.14(.652034^{x})$$

where x measures deviations in five-year units from an origin at 1850.

Succeeding calculations are shown in Table H.

TABLE H

Computation of Ordinates of Trend of Logistic Curve Fitted to Data of Railroad Mileage

(1) Year	(2) x	(3) c^x	(4) bc^x	(5) $\frac{100{,}000{,}000}{y}$ $(a + bc^x)$	(6) y $\left(100{,}000{,}000 \times \frac{1}{(5)}\right)$
1850	0	1.000000	8,921	9,290	10,764
1855	1	.652034	5,817	6,186	16,166
1860	2	.425148	3,793	4,162	24,027
1865	3	.277211	2,473	2,842	35,186
1870	4	.180751	1,613	1,982	50,454
1875	5	.117856	1,051	1,420	70,423
1880	6	.076846	686	1,055	94,787
1885	7	.050106	447	816	122,549
1890	8	.032671	291	660	151,515
1895	9	.021303	190	559	178,891
1900	10	.013890	124	493	202,840
1905	11	.009057	81	450	222,222
1910	12	.005905	53	422	236,967
1915	13	.003850	34	403	248,139
1920	14	.002511	22	391	255,754
1925	15	.001637	15	384	260,417
1930	16	.001067	10	379	263,852
1935	17	.000696	6	375	266,667

The process of calculation is a straightforward one. The reciprocals of the entries in col. (5), multiplied by 100,000,000, yield the desired trend values given in col. (6). These values, with the original series, are shown graphically in Fig. C.

As in the case of the Gompertz curve, the logistic is suitable for measuring the trend of a series that, in its later

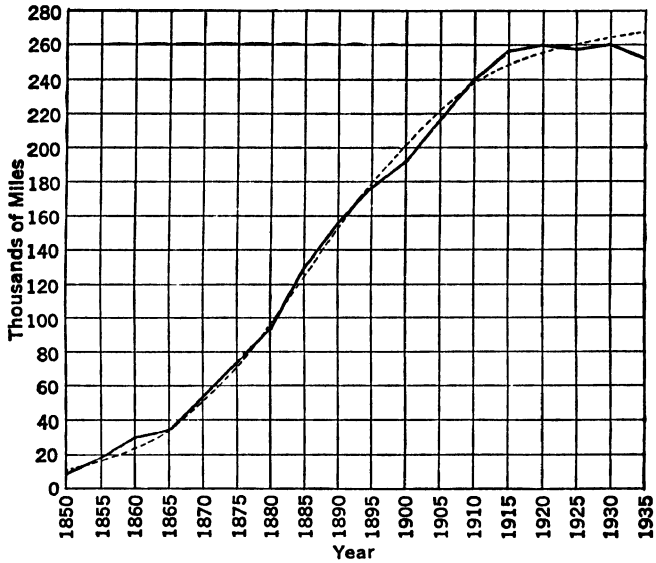


FIG. C. — Railroad Mileage Operated in the United States, by Five-Year Intervals, 1850–1935, with Logistic Trend

stages, is growing by decreasing increments. The curve resembles an elongated S rising from a lower asymptote of zero to an upper limit indicated by the constant a. Since a in this case refers to an equation in which the dependent variable is $\frac{100,000,000}{y}$, the actual asymptote is $\frac{100,000,000}{a}$. From the given value of a, 369.04, we derive 270,973 miles as the upper limit of railroad mileage in the United States. As is clear from the table and chart, the actual values are close to this indicated limit. Barring the possibility of a

fundamental change in relevant conditions, the record and the curve fitted to it indicate that the era of railroad expansion has ended. The extrapolation is, of course, subject to all the reservations that attach to the projection of other curves. There can be no doubt that, within the limits of the observations, the logistic curve gives an excellent representation of the actual history of railroad operation in the United States.

APPENDIX E

A FURTHER APPLICATION OF VARIANCE ANALYSIS

The possibilities of Fisher's method of variance analysis were far from exhausted by the several examples given in Chapter XV. We here supplement the treatment in that chapter by an additional example, illustrating further tests that may be made with a two-fold principle of classification.

The observations on which this example is based consist of relative numbers, measuring the prices of 670 commodities in February, 1933, with average prices in 1926 taken as 100. February, 1933 marked the low point of the severe price decline that began in 1929. The questions to which our tests are directed relate to the relative severity of the declines occurring among different classes of goods.

The 670 price relatives (obtained from price quotations compiled by the U. S. Bureau of Labor Statistics) may be classified into those relating to perishable goods (505 in number) and those relating to durable goods (165 in number). The classification has economic significance because of differences in the market conditions, on both supply and demand sides, affecting these classes of goods during a major recession. Again, the 670 observations may be broken down into those relating to raw materials (134 in number) and those relating to manufactured goods (536 in number). Applying the two principles of classification jointly we obtain 4 subgroups, perishable raw materials (101 in number), perishable manufactured goods (404 in number), durable raw materials (33 in number) and durable manufactured goods (132 in number). It is to be noted that the ratio of the number of perishable raw materials to the number of perishable

manufactured goods, 101:404, is the same as the ratio of the number of durable raw materials to the number of durable manufactured goods, 33:132. It is a necessary condition of the procedure here discussed that the frequencies in the several sub-groups be proportional.

Various questions relating to the significance of these principles of classification may be answered with reference to the summary figures given in Table I.

TABLE I

Measurements Relating to the Analysis of the Relative Prices of 670 Commodities for February, 1933

(1926 = 100)

1 *Perishable raw materials*	2 *Perishable manufactured goods*	I *All perishable goods*
$N_1 = 101$	$N_2 = 404$	$N_p = 505$
$M_1 = 41.663366$	$M_2 = 62.329208$	$M_p = 58.196040$
$\Sigma d^2 = 31{,}118.56$	$\Sigma d^2 = 187{,}414.21$	$\Sigma d^2 = 253{,}040.57$
3 *Durable raw materials*	4 *Durable manufactured goods*	II *All durable goods*
$N_3 = 33$	$N_4 = 132$	$N_d = 165$
$M_3 = 65.060606$	$M_4 = 75.719697$	$M_d = 73.587879$
$\Sigma d^2 = 12{,}217.88$	$\Sigma d^2 = 31{,}308.63$	$\Sigma d^2 = 46{,}525.97$
A *All raw materials*	B *All manufactured goods*	*All commodities*
$N_r = 134$	$N_m = 536$	$N = 670$
$M_r = 47.425373$	$M_m = 65.626866$	$M = 61.986567$
$\Sigma d^2 = 56{,}952.76$	$\Sigma d^2 = 236{,}562.35$	$\Sigma d^2 = 329{,}029.89$

The entries relating to each group and sub-group define the number of commodities included, the mean value of the price relatives for February, 1933, and the sum of the squares of the deviations of the observations in that group from the mean of that group. Thus for perishable raw ma-

terials the mean is 41.663366 (indicating an average price decline of 58.34 per cent) and the sum of the squares of the deviations of the 101 observations in this group from 41.663366 is 31,118.56. For all commodities the mean is 61.986567, and the sum of the squares of the deviations of the individual items from this mean is 329,029.89.

A TEST OF THE PERISHABLE-DURABLE PRINCIPLE OF CLASSIFICATION

We may first ask whether the application of the two basic principles of classification, considered separately, gives groups showing significant differences in their price changes from 1926 to February, 1933. Examining the results of the perishable-durable distinction, we note that durable goods, with an average of 73.587879, show smaller price declines than perishable goods, for which the average is 58.196040. (Six decimal places are retained in the averages because these figures enter into later calculations.) Is the difference significant, or may it be attributed to chance? A test of the type discussed in Chapter XV provides an answer to this question. For the application of the test we must divide the total variability, 329,029.89, into a portion unaffected by perishable-durable differences and a portion that may be attributed to the play of forces directly related to this distinction.

The first of these portions, measuring the variability within classes, is derived directly from the figures in Table I.

Variability within perishable group = Σd^2 for that group	= 253,040.57
Variability within durable group = Σd^2 for that group	= 46,525.97
Total variability within classes	299,566.54

In deriving a measure of the variability between classes we take the deviation of each class mean from the mean of all the observations, square this, and weight by the number

of observations in that class. Thus

Σd^2 between perishable-durable classes
$= [(61.986567 - 58.196040)^2 \times 505] + [(61.986567 - 73.587879)^2 \times 165] = 29{,}463.31.$

A test of the significance of this classification reduces to the question whether the variability between classes is significantly greater than the variability within classes, when account has been taken of the number of degrees of freedom present in the two measures of variability. The appropriate z-test is shown below.

Nature of variability	*Degrees of freedom* n	*Sum of squares*	*Variance* σ^2	$Log_e \sigma^2$
Between classes	1	29,463.31	29,463.31	10.290900
Within classes	668	299,566.54	448.45	6.105804
	669	329,029.85	Diff. =	4.185096
			z =	2.09

For $n_1 = 1$ and $n_2 = 668$ the 1 per cent value of z is approximately .95; the present value is materially greater than this. The variance between classes is significantly greater than the variance within classes. The results are not consistent with the hypothesis that the true value of z is zero. There is a significant difference between the February, 1933, price relatives of perishable and durable goods, on the 1926 base. This principle of classification is a significant one, with reference to this aspect of price behavior.

A TEST OF THE RAW-MANUFACTURED PRINCIPLE OF CLASSIFICATION

The test of the other main principle of classification follows exactly the same lines. The total variability, 329,029.89, is broken into a portion measuring variability within classes (293,515.11), with 668 degrees of freedom, and a portion measuring variability between the raw-manufactured classes (35,514.75) with 1 degree of freedom. The value of z is 2.20; the corresponding 1 per cent value of z is .95. This

principle of classification, also, is significant. Raw and manufactured goods differed significantly in degree of price change between 1926 and February, 1933.

A TEST OF THE RESULTS OBTAINED FROM THE JOINT APPLICATION OF THE PERISHABLE-DURABLE AND RAW-MANUFACTURED PRINCIPLES OF CLASSIFICATION

The application of the two principles of classification discussed above yields the 4 cells shown in Table I. We may ask whether the four groups thus distinguished — perishable raw materials, perishable manufactured goods, durable raw materials, and durable manufactured goods — are significantly different, judged with reference to the present observations. The two essential elements of the total variability are derived from the figures in Table I in the manner indicated below.

Variability within perishable raw materials group =	31,118.56
Variability within perishable manufactures group =	187,414.21
Variability within durable raw materials group =	12,217.88
Variability within durable manufactures group =	31,308.63
Total variability within cells	262,059.28

This sum furnishes the yardstick that is used in the tests that follow. It is clear that it represents the action of forces other than those related to relative durability, or to degree of fabrication. For its four elements measure variability among commodities that are alike in respect of durability and alike in respect of degree of fabrication.[1] This sum is a measure of the strength of the forces we lump together as chance, which here means all factors affecting our observations other than those related to the relative durability of commodities or to degree of fabrication of commodities.

[1] This statement may be accepted as accurate for the purpose of the present demonstration. Actually, of course, the distinctions between perishable and durable commodities and between raw and manufactured goods are not clear-cut and definitive.

A measure of variability between cells is derived as in the previous examples.

$$\begin{aligned}\Sigma d^2 \text{ between cells} &= [(61.986567 - 41.663366)^2 \times 101] \\ &\quad + [(61.986567 - 62.329208)^2 \times 404] \\ &\quad + [(61.986567 - 65.060606)^2 \times 33] \\ &\quad + [(61.986567 - 75.719697)^2 \times 132] \\ &= 66{,}970.60.\end{aligned}$$

The test of significance takes the following form.

Nature of variability	*Degrees of freedom* n	*Sum of squares*	*Variance* σ^2	$Log_e \sigma^2$
Between cells	3	66,970.60	22,323.53	10.013395
Within cells	666	262,059.28	393.48	5.975035
	669	329,029.88		Diff. = 4.038360
				z = 2.02

For $n_1 = 3$, $n_2 = 666$ the 1 per cent value of z is approximately .67. The present value materially exceeds this. The conclusion is clear that the joint application of the two principles of classification yields sub-groups which differed significantly in their price movements between 1926 and February, 1933.

FURTHER TESTS OF THE MAIN PRINCIPLES OF CLASSIFICATION

The test applied in the preceding section does not bring out the most significant uses of a two-fold principle of classification. In treating the four cells as we have, we have not made full use of the information we possess about them. The variance between cells, measured by the sum 66,970.60, with 3 degrees of freedom, represents the combined influence of forces related to the perishable-durable principle of classification, to the raw-manufactured principle, and to the interaction among forces related to these two principles. We may apply more refined tests, and obtain more accurate information about the differential price behavior of commodities of different types, by distinguishing the components of the variance between cells. This is done in Table J,

which presents a complete breakdown of the total variance of the observations with which we are working.

TABLE J

Components of Variance among Observations Relating to Commodity Price Movements, 1926 — February, 1933

(1926 = 100)

Nature of variability	*Degrees of freedom*	*Sum of squares*	*Variance* σ^2
Between perishable-durable classes	1	29,463.31	29,463.31
Between raw-manufactured classes	1	35,514.75	35,514.75
Interaction (residual variability between cells)	1	1,992.54	1,992.54
Within cells ("experimental error")	666	262,059.28	393.48
	669	329,029.88	

Having these components we may test with greater accuracy than on pages 683 and 684 the significance of the two main principles of classification. For we now have a better yardstick, a better measure of the magnitude of variations due to the play of "chance." The variability within cells (variance = 393.48) is a better criterion of the magnitude of sampling errors than is the variability within the perishable and durable classes (variance = 448.45) or the variability within the raw and manufactured classes (variance = 439.39). For the variance within the four cells is free of the influence of forces connected with either of the specified principles of classification.

This more accurate test of the perishable-durable principle of classification is applied by the customary method.

Nature of variability	*Degrees of freedom*	*Variance* σ^2	$Log_e \sigma^2$
Between perishable-durable classes	1	29,463.31	10.290900
Within cells	666	393.48	5.975035
			Diff. = 4.315865
			z = 2.16

The 1 per cent value of z is approximately .95, and the above result is clearly significant. The application of the perishable-durable principle of classification, under the conditions represented in Table I, yields classes of commodities that differed significantly in their price changes between 1926 and February, 1933. It is important to note that raw and manufactured goods are present in the perishable and durable groups in precisely the same proportions. One fifth of the commodities in each group are raw materials and four fifths are manufactured goods. Thus behavior peculiar to raw materials may be expected to influence the two groups in precisely the same degree; the same is true of behavior peculiar to goods in the manufactured state.[1] It is necessary, for this reason, that the frequencies in the several classes be proportional in the application of the tests here discussed, when two principles of classification are jointly employed.[2]

A test of the significance of the raw-manufactured principle of classification may be applied in the same way. The variance within cells is employed as yardstick, as in the preceding example. Here, also, proportionality is necessary, with raw and manufactured goods being divided in the same proportions into perishable and durable sub-groups. The test reveals a significant difference in price behavior between raw and manufactured commodities.

A TEST OF THE INTERACTION

Not all the variability between cells is explained by the two major classifications we have just discussed. The residual variability between cells, or the *interaction*, amounts to 1,992.54, in terms of squared deviations (see Table J).

[1] See below, however, for a test of the significance of the *interaction*.

[2] For a discussion of procedures appropriate to cases in which cell frequencies are not proportional see

Yates, F. *Journal of Agricultural Science*, Vol. 23, 108 (1933).

Snedecor, G. W. and Cox, G. M. *Iowa Agricultural Experiment Station Bulletin* 180 (1935).

This may be derived readily by subtracting from the total variability between cells (66,970.60) the sum of the variability between perishable-durable classes (29,463.31) and the variability between raw-manufactured classes (35,514.75). The number of degrees of freedom in the *interaction* may be determined by the same process of subtraction. In the present instance it is 1.

This residual variability may represent "experimental error," the play of the same chance forces that are measured by the variability within cells. The residual variability was used, in the last example cited in Chapter XV, as a yardstick defining the magnitude of fluctuations due to chance. It is proper to assume that this is the case when the two major principles of classification are quite independent of one another. But if these principles are correlated, the residual variability reflects the *interaction* of the two principles of classification — the differential behavior of given classes of goods under the influence of forces related to the other principle of classification. Thus it may be that the difference between raw perishable and manufactured perishable goods is not the same as the difference between raw durable and manufactured durable goods. The process of fabrication applied to perishable goods may produce results (in the form of price behavior) different from those produced when the process of fabrication is applied to durable goods. Perishable and durable goods may respond differently, as regards their price behavior, to the influence of fabrication. Such differential behavior of categories of goods under the influence of the same treatment (i.e., fabrication) is measured by the *interaction*.

If there is no such differential behavior, in a given experiment, the residual variability between cells will be of the same order of magnitude as the variability within cells, when account is taken of number of degrees of freedom. A test is applied on page 690.

If we judge this result with reference to the 1 per cent

Nature of variability	*Degrees of freedom*	*Sum of squares*	*Variance* σ^2	$Log_e \sigma^2$
Interaction (residual variability between cells)	1	1,922.54	1,922.54	7.561429
Within cells	666	262,059.28	393.48	5.975035
			Diff. =	1.586394
			z =	.79

value of z (.95), we would conclude that the residual variability between cells is attributable to the play of chance rather than to any true *interaction*. For although the residual variability is greater than the variance within cells which we use as yardstick, the excess is not clearly too great to be attributed to chance. Reference to the 5 per cent value of z (.675, for $n_1 = 1$, $n_2 = 666$) throws more light on the situation. Less frequently than 5 times out of 100 would the play of chance alone give us a measure of residual variability as great as that here obtained. For the z of .79 is greater than the 5 per cent value, .675. In such a case as this, where P falls between .01 and .05, the evidence is not conclusive. There is, however, a strong indication that perishable and durable goods respond differently, in their price behavior, to the process of fabrication. Reference to Table I will show that among both perishable and durable goods fabrication appears to have reduced susceptibility to price decline under the force of business recession. M_2 is distinctly greater than M_1, and M_4 is greater than M_3. But the influence of fabrication was apparently greater among perishable than among durable goods.[1] Our test shows that the degree of difference between the two reductions (i.e., reductions in degree of price decline) is almost too great to be attributed to chance. The evidence of differential behavior is strong enough to justify further investigation.

[1] The statistical evidence does not, of course, yield information as to the nature of the causal relations involved. The test here applied, if positive, reveals the presence of interaction, but does not show how the forces involved interact to bring about the observed differential behavior. The text is to be read with this qualification in mind.

APPENDIX F

GLOSSARY OF SYMBOLS

The following are the more important symbols employed in the preceding pages. Those of which limited use is made, for special purposes, are not here included. A given symbol is sometimes called upon to serve different purposes, but the precise meaning should be clear from the context.

1. General symbols for variables and constants:

   x: a variable quantity.

   y: a variable quantity.

   In general, any letter near the end of the alphabet may be employed to represent a variable quantity. Different variable quantities may be represented by the use of a single symbol, with different subscripts, as X_1, X_2, X_3, or W_1, W_2, W_3. [A distinction is later drawn (cf. Symbols employed in the measurement of relationship) between capital letters and small letters, as used to represent variable quantities.]

   a: a constant (i.e., a quantity the value of which does not change in the given discussion). In general, any letter near the beginning of the alphabet may be used to represent a constant.

2. Symbols employed in the analysis and description of the frequency distribution:

   m: the value of an individual observation; the value of the mid-point of a class. (The symbols a_1, a_2, a_3 are sometimes employed to represent different observations in a series.)

   f: the number of observations in a given class; the frequency of a given class.

   i: the class-interval.

   l: the lower limit of a class.

   N: the total number of cases in a given series or frequency distribution.

d: the deviation of a given observation from an average; usually, a deviation from the arithmetic mean. When written with a subscript, as d_x or d_y, it refers to a deviation from the arithmetic mean of the variable represented by the subscript. The symbol d is sometimes used to designate the difference between mean and mode.

d': the deviation of a given observation from an arbitrary origin, or assumed mean.

c: the difference between an arbitrary origin, or assumed mean, and the true mean (in terms of the symbols explained below, $c = M - M'$).

Σ (Sigma): the symbol for the process of summation. Thus Σd means the sum of all the deviations.

w_1, w_2, w_3: weights attached to a series of measures being averaged. (Not to be confused with similar symbols used to represent different variable quantities.)

y_0: the maximum ordinate of a frequency curve.

Symbols for averages, quartiles, etc.:

M: the arithmetic mean.

$Md.$: the median.

$Mo.$: the mode.

M_g: the geometric mean.

H: the harmonic mean.

M': the value of an assumed arithmetic mean.

Q_1: the first or lower quartile.

Q_2: the second quartile or median.

Q_3: the third or upper quartile.

K: the value of a point midway between the first and third quartiles.

D_3: the third decile.

Symbols for measures of variation and skewness:

$M.D.$: the mean deviation.

σ: the standard deviation; the root-mean-square deviation about the arithmetic mean.

σ': the standard deviation of proportions, or relative frequencies.

s_a: the root-mean-square deviation about an origin other than the arithmetic mean.

$P.E.$: the probable error.

$Q.D.$: the quartile deviation.

q_1: the difference between the median and the lower quartile ($Md. - Q_1$).

q_2: the difference between the upper quartile and the median ($Q_3 - Md.$).

V: the coefficient of variation.

sk: a measure of skewness.

χ (Chi): a measure of skewness based upon the criteria β_1 and β_2.

Symbols for moments and criteria of curve type.

ν_1, ν_2, ν_3, etc.: moments of a frequency distribution about an arbitrary origin.

π_1, π_2, π_3, etc.: uncorrected moments of a frequency distribution about the arithmetic mean.

μ_1, μ_2, μ_3, etc.: moments of a frequency distribution about the arithmetic mean after the application of Sheppard's corrections.

β_1: $\frac{\mu_3^2}{\mu_2^3}$.

β_2: $\frac{\mu_4}{\mu_2^2}$.

κ_2: A criterion of curve type based on β_1 and β_2.

3. Symbols relating to index numbers.

p_0': price of a given commodity at time "0" (the base period).

q_0': quantity of same commodity at time "0".

p_1': price of same commodity at time "1".

q_1': quantity of same commodity at time "1".

p_0'': price of a second commodity at time "0".

q_0'': quantity of second commodity at time "0".

p_1'': price of second commodity at time "1".

q_1'': quantity of second commodity at time "1".

$\frac{p_1'}{p_0'}$: a price relative (relation of price of a given commodity at time "1" to price of same commodity at time "0").

$\frac{q_1'}{q_0'}$: a quantity relative.

P_0: price level at time "0".

P_1: price level at time "1".

4. Symbols employed in the measurement of relationship.

X: an observed value of a variable quantity.

Y: an observed value of a variable quantity. (The observed values of different variables may be represented also by the symbols X_1, X_2, X_3, or W_1, W_2, W_3.)

$\overline{X}$: the arithmetic mean of a number of observed values of the variable X. A similar symbol may be employed for other variables. (In one demonstration in the preceding pages, relating to multiple correlation, the symbols A_1, A_2, A_3 . . . are used to represent the arithmetic means of the variables X_1, X_2, X_3 . . . . The symbols M_x and M_y are occasionally employed to designate the arithmetic means of different variables.)

x: value of a variable quantity expressed as a deviation from the arithmetic mean of all the observed values. The symbol y and the symbols x_1, x_2, x_3 . . . are similarly employed with respect to variables represented, as to original observations, by the symbols Y, X_1, X_2, X_3 . . . .

X': a value of a variable quantity expressed as a deviation, in class-interval units, from an arbitrary origin. The symbol Y' has a similar meaning.

X'': a value of a variable quantity expressed as a deviation, in original units, from an arbitrary origin. The symbol Y'' has a similar meaning.

Y_c: the computed or estimated value of a variable, as determined from an equation of average relationship; the symbol y_c may be employed for such a computed value, expressed as a deviation from the mean.

p: the mean product of two variables when expressed as deviations from their respective arithmetic means, i.e., $p = \frac{\Sigma(xy)}{N}$. When written with subscripts, as p_{12}, the latter relate to the variables in question, as x_1, x_2.

p': the mean product of two variables when expressed as deviations from assumed arithmetic means.

r: the Pearsonian coefficient of correlation. When written with subscripts, the latter indicate the variables to which the coefficient relates. Thus r_{yx} refers to the variables y and x, and r_{12} refers to the variables x_1 and x_2.

ρ (rho): a general index of correlation. Subscripts should be

employed to indicate the variables to which the measure relates, as ρ_{yx}, ρ_{xy}, $\rho_{\log yx}$, $\rho_{\log y \log x}$, $\rho_{\frac{1}{y}x}$, etc. (In each case the first subscript relates to the dependent variable.)

$\bar{\rho}$: a corrected index of correlation.

d: the deviation of a given observation from a fitted curve; the difference between an observed and a corresponding computed value of a variable.

v: a residual; identical in meaning with d, as given above.

S: the root-mean-square deviation about a fitted curve; the standard error of estimate. This measure should be written with a subscript to indicate the variable to which it applies, as S_y, S_x, $S_{\log y}$ (the standard error of estimate in terms of logarithms), S_r (the standard error of estimate in terms of ratios), $S_{\frac{1}{y}}$ (the standard error of estimate in terms of reciprocals).

$\bar{S}$: a corrected standard error of estimate.

ρ_r: the coefficient of rank correlation.

η (eta): the correlation ratio. Subscripts should be employed to represent the variables to which the measure relates, as η_{yx} or η_{xy}. The first subscript in each case relates to the dependent variable.

$\bar{\eta}$: a corrected correlation ratio.

σ_{ay}: the root-mean-square deviation about a line through the means of the columns of a correlation table; the standard deviation of the y-arrays about their respective means. The symbol σ_{ax} has the same meaning with respect to the rows of a correlation table, or the x-arrays.

σ_{my}: the standard deviation of the means of the columns of a correlation table about the mean of all the y's, the mean of each column being weighted by the number of items in that column. The symbol σ_{mx} has the same meaning with respect to the means of the rows.

ζ (zeta): the test for linearity of regression ($\zeta = \eta^2 - r^2$).

m: the number of arrays employed in the computation of a given correlation ratio; also, the number of constants in the equation defining a curvilinear or multiple regression.

b: the coefficient of regression; the slope of a line of regression. When written with subscripts, the latter relate

to the variables in question, as b_{yx}, b_{12} (for the variables x_1, x_2). The first subscript relates to the dependent variable in each case; b_{yx} is the coefficient of regression of y on x and b_{xy} is the coefficient of regression of x on y.

z: a logarithmic transformation of the coefficient of correlation. $z = \frac{1}{2}\{\log_e (1 + r) - \log_e (1 - r)\}$.

$R_{1.234}$: the coefficient of multiple correlation between a dependent variable, x_1, and a combination of independent variables, x_2, x_3, and x_4. The order may be changed, but the primary subscript always relates to the dependent variable.

$\bar{R}_{1.234}$: a corrected coefficient of multiple correlation.

$r_{12.34}$: the coefficient of partial or net correlation between the variables x_1 and x_2, when the variables x_3 and x_4 are held constant. The order of subscripts is changed for a different combination of variables, the two primary subscripts always relating to the variables between which the net correlation is being measured.

$b_{12.34}$: the coefficient of net regression between the variables x_1 and x_2, the former being dependent, when the variables x_3 and x_4 are also taken account of in the estimating equation; the weight given to x_2 in estimating x_1, when the estimate is also based upon values of x_3 and x_4. The order of subscripts is changed for a different combination of variables.

$S_{1.234}$: the root-mean-square deviation about a line describing the relationship between a dependent variable, x_1, and a series of independent variables, x_2, x_3, and x_4; the standard error of estimate of x_1 under these conditions.

$\sigma_{1.234}$: the standard deviation of the fourth order; identical with $S_{1.234}$.

$\beta_{12.34}$: a coefficient of partial regression in an equation relating to variables expressed in standard deviation units.

(In the seven measures immediately above, the number of subscripts corresponds to the number of variables included in a given study. For the sake of simplicity, only four variables have been assumed.)

5. Symbols employed in the measurement of errors.

σ: the standard deviation of a parent population.

σ_M or $\sigma_{\bar{x}}$: the standard error of a mean, derived from a knowledge of the σ of the population.

s: the standard deviation of a sample.

s_M or $s_{\bar{x}}$: the estimated standard error of a mean, in the derivation of which s is used as an approximation to σ.

T: the deviation of a given statistical measurement from the mean of a normal distribution, expressed in units of the standard deviation of that distribution; a normal deviate.

t: the deviation of a given statistical measurement from a hypothetical value, expressed in units of the estimated standard error of the measurement in question.

σ_{M_s}: the standard error of the mean of a stratified sample.

D: a difference between two means.

σ_D: the standard error of the difference between two means.

D_p: a difference between two percentages.

σ_{D_p}: The standard error of the difference between two percentages.

D_z: the difference between two logarithmic transformations of the coefficient of correlation.

σ_{D_z}: the standard error of D_z.

σ, with any subscript, is used to represent the standard error of the measure to which the subscript relates.

P.E. with any subscript is used to represent the probable error of the measure to which the subscript relates (P.E. $= .67449\sigma$).

$\sigma_{b_1-b_2}$: the standard error of the difference between two coefficients of regression.

6. Symbols employed in the analysis of variance.

z: the difference between the natural logarithms of two standard deviations.

σ_z: the standard error of z.

n_1: the number of degrees of freedom in the larger of two variances being compared.

n_2: the number of degrees of freedom in the smaller of two variances being compared.

7. Other symbols.

p: the probability of a successful outcome of a given event.

q: the probability of an unsuccessful outcome of a given event.

n: the number of independent events in a given trial.

χ^2: a quantity used in testing hypotheses involving the computation of theoretical frequencies; χ^2 defines the relative magnitude of the differences between observed and theoretical frequencies.

GREEK ALPHABET

Letters	*Names*	*Letters*	*Names*	*Letters*	*Names*
A α	Alpha	I ι	Iota	P ρ	Rho
B β	Beta	K κ	Kappa	Σ σ	Sigma
Γ γ	Gamma	Λ λ	Lambda	T τ	Tau
Δ δ	Delta	M μ	Mu	Υ υ	Upsilon
E ϵ	Epsilon	N ν	Nu	Φ ϕ	Phi
Z ζ	Zeta	Ξ ξ	Xi	X χ	Chi
H η	Eta	O o	Omicron	Ψ ψ	Psi
Θ θ	Theta	Π π	Pi	Ω ω	Omega

Appendix Table I

Areas of the Normal Curve of Error in Terms of Abscissa

$\frac{x}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.00000	.00399	.00798	.01197	.01595	.01994	.02392	.02790	.03188	.03586
0.1	.03983	.04380	.04776	.05172	.05567	.05962	.06356	.06749	.07142	.07535
0.2	.07926	.08317	.08706	.09095	.09483	.09871	.10257	.10642	.11026	.11409
0.3	.11791	.12172	.12552	.12930	.13307	.13683	.14058	.14431	.14803	.15173
0.4	.15542	.15910	.16276	.16640	.17003	.17364	.17724	.18082	.18439	.18793
0.5	.19146	.19497	.19847	.20194	.20540	.20884	.21226	.21566	.21904	.22240
0.6	.22575	.22907	.23237	.23565	.23891	.24215	.24537	.24857	.25175	.25490
0.7	.25804	.26115	.26424	.26730	.27035	.27337	.27637	.27935	.28230	.28524
0.8	.28814	.29103	.29389	.29673	.29955	.30234	.30511	.30785	.31057	.31327
0.9	.31594	.31859	.32121	.32381	.32639	.32894	.33147	.33398	.33646	.33891
1.0	.34134	.34375	.34614	.34850	.35083	.35314	.35543	.35769	.35993	.36214
1.1	.36433	.36650	.36864	.37076	.37286	.37493	.37698	.37900	.38100	.38298
1.2	.38493	.38686	.38877	.39065	.39251	.39435	.39617	.39796	.39973	.40147
1.3	.40320	.40490	.40658	.40824	.40988	.41149	.41309	.41466	.41621	.41774
1.4	.41924	.42073	.42220	.42364	.42507	.42647	.42786	.42922	.43056	.43189
1.5	.43319	.43448	.43574	.43699	.43822	.43943	.44062	.44179	.44295	.44408
1.6	.44520	.44630	.44738	.44845	.44950	.45053	.45154	.45254	.45352	.45449
1.7	.45543	.45637	.45728	.45818	.45907	.45994	.46080	.46164	.46246	.46327
1.8	.46407	.46485	.46562	.46638	.46712	.46784	.46856	.46926	.46995	.47062
1.9	.47128	.47193	.47257	.47320	.47381	.47441	.47500	.47558	.47615	.47670
2.0	.47725	.47778	.47831	.47882	.47932	.47982	.48030	.48077	.48124	.48169
2.1	.48214	.48257	.48300	.48341	.48382	.48422	.48461	.48500	.48537	.48574
2.2	.48610	.48645	.48679	.48713	.48745	.48778	.48809	.48840	.48870	.48899
2.3	.48928	.48956	.48983	.49010	.49036	.49061	.49086	.49111	.49134	.49158
2.4	.49180	.49202	.49224	.49245	.49266	.49286	.49305	.49324	.49343	.49361
2.5	.49379	.49396	.49413	.49430	.49446	.49461	.49477	.49492	.49506	.49520
2.6	.49534	.49547	.49560	.49573	.49585	.49598	.49609	.49621	.49632	.49643
2.7	.49653	.49664	.49674	.49683	.49693	.49702	.49711	.49720	.49728	.49736
2.8	.49744	.49752	.49760	.49767	.49774	.49781	.49788	.49795	.49801	.49807
2.9	.49813	.49819	.49825	.49831	.49836	.49841	.49846	.49851	.49856	.49861
3.0	.49865	.49869	.49874	.49878	.49882	.49886	.49889	.49893	.49897	.49900
3.1	.49903	.49906	.49910	.49913	.49916	.49918	.49921	.49924	.49926	.49929
3.2	.49931	.49934	.49936	.49938	.49940	.49942	.49944	.49946	.49948	.49950
3.3	.49952	.49953	.49955	.49957	.49958	.49960	.49961	.49962	.49964	.49965
3.4	.49966	.49968	.49969	.49970	.49971	.49972	.49973	.49974	.49975	.49976

Appendix Table II [1]

Table of t

n	$P = .05$	.02	.01
1	12.706	31.821	63.657
2	4.303	6.965	9.925
3	3.182	4.541	5.841
4	2.776	3.747	4.604
5	2.571	3.365	4.032
6	2.447	3.143	3.707
7	2.365	2.998	3.499
8	2.306	2.896	3.355
9	2.262	2.821	3.250
10	2.228	2.764	3.169
11	2.201	2.718	3.106
12	2.179	2.681	3.055
13	2.160	2.650	3.012
14	2.145	2.624	2.977
15	2.131	2.602	2.947
16	2.120	2.583	2.921
17	2.110	2.567	2.898
18	2.101	2.552	2.878
19	2.093	2.539	2.861
20	2.086	2.528	2.845
21	2.080	2.518	2.831
22	2.074	2.508	2.819
23	2.069	2.500	2.807
24	2.064	2.492	2.797
25	2.060	2.485	2.787
26	2.056	2.479	2.779
27	2.052	2.473	2.771
28	2.048	2.467	2.763
29	2.045	2.462	2.756
30	2.042	2.457	2.750
∞	1.95996	2.32634	2.57582

[1] Excerpts from Table IV, R. A. Fisher, *Statistical Methods for Research Workers*. These excerpts are printed here through the courtesy of Dr. Fisher and his publishers, Oliver and Boyd, of Edinburgh.

Appendix Table III [1]

Values of the Correlation Coefficient for Different Levels of Significance

n	$P = .05$	.02	.01
1	.996917	.9995066	.9998766
2	.95000	.98000	.990000
3	.8783	.93433	.95873
4	.8114	.8822	.91720
5	.7545	.8329	.8745
6	.7067	.7887	.8343
7	.6664	.7498	.7977
8	.6319	.7155	.7646
9	.6021	.6851	.7348
10	.5760	.6581	.7079
11	.5529	.6339	.6835
12	.5324	.6120	.6614
13	.5139	.5923	.6411
14	.4973	.5742	.6226
15	.4821	.5577	.6055
16	.4683	.5425	.5897
17	.4555	.5285	.5751
18	.4438	.5155	.5614
19	.4329	.5034	.5487
20	.4227	.4921	.5368
25	.3809	.4451	.4869
30	.3494	.4093	.4487
35	.3246	.3810	.4182
40	.3044	.3578	.3932
45	.2875	.3384	.3721
50	.2732	.3218	.3541
60	.2500	.2948	.3248
70	.2319	.2737	.3017
80	.2172	.2565	.2830
90	.2050	.2422	.2673
100	.1946	.2301	.2540

For a total correlation, n is 2 less than the number of pairs in the sample; for a partial correlation, the number of eliminated variates also should be subtracted.

[1] Excerpts from Table V–A, R. A. Fisher, *Statistical Methods for Research Workers*. These excerpts are printed here through the courtesy of Dr. Fisher and his publishers, Oliver and Boyd, of Edinburgh.

Appendix Table IV

Showing the Relations between r and z for Values of z from 0 to 5[1]

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0100	.0200	.0300	.0400	.0500	.0599	.0699	.0798	.0898
.1	.0997	.1096	.1194	.1293	.1391	.1489	.1587	.1684	.1781	.1878
.2	.1974	.2070	.2165	.2260	.2355	.2449	.2543	.2636	.2729	.2821
.3	.2913	.3004	.3095	.3185	.3275	.3364	.3452	.3540	.3627	.3714
.4	.3800	.3885	.3969	.4053	.4136	.4219	.4301	.4382	.4462	.4542
.5	.4621	.4700	.4777	.4854	.4930	.5005	.5080	.5154	.5227	.5299
.6	.5370	.5441	.5511	.5581	.5649	.5717	.5784	.5850	.5915	.5980
.7	.6044	.6107	.6169	.6231	.6291	.6352	.6411	.6469	.6527	.6584
.8	.6640	.6696	.6751	.6805	.6858	.6911	.6963	.7014	.7064	.7114
.9	.7163	.7211	.7259	.7306	.7352	.7398	.7443	.7487	.7531	.7574
1.0	.7616	.7658	.7699	.7739	.7779	.7818	.7857	.7895	.7932	.7969
1.1	.8005	.8041	.8076	.8110	.8144	.8178	.8210	.8243	.8275	.8306
1.2	.8337	.8367	.8397	.8426	.8455	.8483	.8511	.8538	.8565	.8591
1.3	.8617	.8643	.8668	.8693	.8717	.8741	.8764	.8787	.8810	.8832
1.4	.8854	.8875	.8896	.8917	.8937	.8957	.8977	.8996	.9015	.9033
1.5	.9052	.9069	.9087	.9104	.9121	.9138	.9154	.9170	.9186	.9202
1.6	.9217	.9232	.9246	.9261	.9275	.9289	.9302	.9316	.9329	.9342
1.7	.9354	.9367	.9379	.9391	.9402	.9414	.9425	.9436	.9447	.9458
1.8	.9468	.9478	.9498	.9488	.9508	.9518	.9527	.9536	.9545	.9554
1.9	.9562	.9571	.9579	.9587	.9595	.9603	.9611	.9619	.9626	.9633
2.0	.9640	.9647	.9654	.9661	.9668	.9674	.9680	.9687	.9693	.9699
2.1	.9705	.9710	.9716	.9722	.9727	.9732	.9738	.9743	.9748	.9753
2.2	.9757	.9762	.9767	.9771	.9776	.9780	.9785	.9789	.9793	.9797
2.3	.9801	.9805	.9809	.9812	.9816	.9820	.9823	.9827	.9830	.9834
2.4	.9837	.9840	.9843	.9846	.9849	.9852	.9855	.9858	.9861	.9863
2.5	.9866	.9869	.9871	.9874	.9876	.9879	.9881	.9884	.9886	.9888
2.6	.9890	.9892	.9895	.9897	.9899	.9901	.9903	.9905	.9906	.9908
2.7	.9910	.9912	.9914	.9915	.9917	.9919	.9920	.9922	.9923	.9925
2.8	.9926	.9928	.9929	.9931	.9932	.9933	.9935	.9936	.9937	.9938
2.9	.9940	.9941	.9942	.9943	.9944	.9945	.9946	.9947	.9949	.9950
3.0	.9951									
4.0	.9993									
5.0	.9999									

[1] The figures in the body of the table are values of r corresponding to z-values read from the scales on the left and top of the table.

Appendix Table V [1]

Table of χ^2

n	$P = .99$	.95	.50	.10	.05	.02	.01
1	.000157	.00393	.455	2.706	3.841	5.412	6.635
2	.0201	.103	1.386	4.605	5.991	7.824	9.210
3	.115	.352	2.366	6.251	7.815	9.837	11.341
4	.297	.711	3.357	7.779	9.488	11.668	13.277
5	.554	1.145	4.351	9.236	11.070	13.388	15.086
6	.872	1.635	5.348	10.645	12.592	15.033	16.812
7	1.239	2.167	6.346	12.017	14.067	16.622	18.475
8	1.646	2.733	7.344	13.362	15.507	18.168	20.090
9	2.088	3.325	8.343	14.684	16.919	19.679	21.666
10	2.558	3.940	9.342	15.987	18.307	21.161	23.209
11	3.053	4.575	10.341	17.275	19.675	22.618	24.725
12	3.571	5.226	11.340	18.549	21.026	24.054	26.217
13	4.107	5.892	12.340	19.812	22.362	25.472	27.688
14	4.660	6.571	13.339	21.064	23.685	26.873	29.141
15	5.229	7.261	14.339	22.307	24.996	28.259	30.578
16	5.812	7.962	15.338	23.542	26.296	29.633	32.000
17	6.408	8.672	16.338	24.769	27.587	30.995	33.409
18	7.015	9.390	17.338	25.989	28.869	32.346	34.805
19	7.633	10.117	18.338	27.204	30.144	33.687	36.191
20	8.260	10.851	19.337	28.412	31.410	35.020	37.566
21	8.897	11.591	20.337	29.615	32.671	36.343	38.932
22	9.542	12.338	21.337	30.813	33.924	37.659	40.289
23	10.196	13.091	22.337	32.007	35.172	38.968	41.638
24	10.856	13.848	23.337	33.196	36.415	40.270	42.980
25	11.524	14.611	24.337	34.382	37.652	41.566	44.314
26	12.198	15.379	25.336	35.563	38.885	42.856	45.642
27	12.879	16.151	26.336	36.741	40.113	44.140	46.963
28	13.565	16.928	27.336	37.916	41.337	45.419	48.278
29	14.256	17.708	28.336	39.087	42.557	46.693	49.588
30	14.953	18.493	29.336	40.256	43.773	47.962	50.892

For larger values of n, the expression $\sqrt{2X^2} - \sqrt{2n - 1}$ may be used as a normal deviate with unit standard error.

[1] Excerpts from Table III, R. A. Fisher, *Statistical Methods for Research Workers*. These excerpts are printed here through the courtesy of Dr. Fisher and his publishers, Oliver and Boyd, of Edinburgh.

APPENDIX TABLE VI [1]

1 Per Cent Points of the Distribution of z

Values of n_2	Values of n_1 = 1.	2.	3.	4.	5.	6.	8.	12.	24.	∞.
1	4.1535	4.2585	4 2974	4.3175	4.3297	4.3379	4.3482	4.3585	4.3689	4.3794
2	2.2950	2.2976	2.2984	2.2988	2.2991	2.2992	2.2994	2.2997	2.2999	2.3001
3	1.7649	1.7140	1.6915	1.6786	1.6703	1.6645	1.6569	1.6489	1.6404	1 6314
4	1.5270	1.4452	1.4075	1.3856	1.3711	1.3609	1.3473	1.3327	1.3170	1.3000
5	1.3943	1.2929	1.2449	1 2164	1 1974	1.1838	1.1644	1.1457	1.1239	1.0997
6	1.3103	1.1955	1.1401	1.1068	1.0843	1.0680	1.0460	1.0218	.9948	.9643
7	1.2526	1.1281	1.0672	1.0300	1.0048	.9864	.9614	.9335	.9020	.8658
8	1.2106	1.0787	1.0135	.9734	.9459	.9259	.8983	.8673	.8319	.7904
9	1.1786	1.0411	.9724	.9299	.9006	.8791	.8494	.8157	.7769	.7305
10	1.1535	1.0114	.9399	.8954	.8646	.8419	.8104	.7744	.7324	.6816
11	1.1333	.9874	.9136	.8674	.8354	.8116	.7785	.7405	.6958	.6408
12	1.1166	.9677	.8919	.8443	.8111	.7864	.7520	.7122	.6649	.6061
13	1.1027	.9511	.8737	.8248	.7907	.7652	.7295	.6882	.6386	.5761
14	1.0909	.9370	.8581	.8082	.7732	.7471	.7103	.6675	.6159	.5500
15	1 0807	.9249	.8448	.7939	.7582	.7314	.6937	.6496	.5961	.5269
16	1.0719	.9144	.8331	.7814	.7450	.7177	.6791	.6339	.5786	.5064
17	1.0641	.9051	.8229	.7705	.7335	.7057	.6663	.6199	.5630	.4879
18	1 0572	.8970	.8138	.7607	.7232	.6950	.6549	.6075	.5516	.4712
19	1.0511	.8897	.8057	.7521	.7140	.6854	.6447	.5964	.5366	.4560
20	1.0457	.8831	.7985	.7443	.7058	.6768	.6355	.5864	.5253	.4421
21	1.0408	.8772	.7920	.7372	.6984	.6690	.6272	.5773	.5150	.4294
22	1.0363	.8719	.7860	.7309	.6916	.6620	.6196	.5691	.5056	.4176
23	1.0322	.8670	.7806	.7251	.6855	.6555	.6127	.5615	.4969	.4068
24	1.0285	.8626	.7757	.7197	.6799	.6496	.6064	.5545	.4890	.3967
25	1 0251	.8585	.7712	.7148	.6747	.6442	.6006	.5481	.4816	.3872
26	1.0220	.8548	.7670	.7103	.6699	.6392	.5952	.5422	.4748	.3784
27	1.0191	.8513	.7631	.7062	.6655	.6346	.5902	.5367	.4685	.3701
28	1.0164	.8481	.7595	.7023	.6614	.6303	.5856	.5316	.4626	.3624
29	1.0139	.8451	.7562	.6987	.6576	.6263	.5813	.5269	.4570	.3550
30	1.0116	.8423	.7531	.6954	.6540	.6226	.5773	.5224	.4519	.3481
60	.9784	.8025	.7086	.6472	.6028	.5687	.5189	.4574	.3746	.2352
∞	.9462	.7636	.6651	.5999	.5522	.5152	4604	.3908	.2913	0

[1] From Table VI, R. A. Fisher, *Statistical Methods for Research Workers.* This table is printed here through the courtesy of Dr. Fisher and his publishers, Oliver and Boyd, of Edinburgh.

Appendix Table VII [1]

5 Per Cent Points of the Distribution of z

Values of n_2	Values of n_1									
	1.	2.	3.	4.	5.	6.	8.	12.	24.	∞.
1	2.5421	2.6479	2.6870	2.7071	2.7194	2.7276	2.7380	2 7484	2.7588	2.7693
2	1.4592	1.4722	1.4765	1.4787	1.4800	1.4808	1.4819	1.4830	1.4840	1.4851
3	1.1577	1.1284	1.1137	1 1051	1.0994	1.0953	1.0899	1.0842	1.0781	1 0716
4	1.0212	.9690	.9429	.9272	.9168	.9093	.8993	.8885	.8767	.8639
5	.9441	.8777	.8441	.8236	.8097	.7997	.7862	.7714	.7550	.7368
6	.8948	.8188	.7798	.7558	.7394	.7274	.7112	.6931	.6729	6499
7	.8606	.7777	.7347	.7080	.6896	.6761	.6576	.6369	.6134	.5862
8	.8355	.7475	.7014	.6725	.6525	.6378	.6175	.5945	.5682	5371
9	.8163	.7242	.6757	.6450	.6238	.6080	.5862	.5613	.5324	.4979
10	.8012	.7058	.6553	.6232	.6009	.5843	.5611	.5346	.5035	.4657
11	.7889	.6909	.6387	.6055	.5822	.5648	.5406	.5126	.4795	.4387
12	.7788	.6786	.6250	.5907	.5666	.5487	.5234	.4941	.4592	.4156
13	.7703	.6682	.6134	.5783	.5535	.5350	.5089	.4785	.4419	.3957
14	.7630	.6594	.6036	.5677	.5423	.5233	.4964	.4649	.4269	.3782
15	.7568	.6518	.5950	.5585	.5326	.5131	.4855	.4532	.4138	.3628
16	.7514	.6451	.5876	.5505	.5241	.5042	.4760	.4428	.4022	.3490
17	.7466	.6393	.5811	.5434	.5166	.4964	.4676	.4337	.3919	.3366
18	.7424	.6341	.5753	.5371	.5099	.4894	.4602	.4255	.3827	.3253
19	.7386	.6295	.5701	.5315	.5040	.4832	.4535	.4182	.3743	.3151
20	.7352	.6254	.5654	.5265	.4986	.4776	.4474	.4116	.3668	.3057
21	.7322	.6216	.5612	.5219	.4938	.4725	.4420	.4055	.3599	.2971
22	.7294	.6182	.5574	.5178	.4894	.4679	.4370	.4001	.3536	.2892
23	.7269	.6151	.5540	.5140	.4854	.4636	.4325	.3950	.3478	.2818
24	.7246	.6123	.5508	.5106	.4817	.4598	.4283	.3904	.3425	.2749
25	.7225	.6097	.5478	.5074	.4783	.4562	.4244	.3862	.3376	.2685
26	.7205	.6073	.5451	.5045	.4752	.4529	.4209	.3823	.3330	.2625
27	.7187	.6051	.5427	.5017	.4723	.4499	.4176	.3786	.3287	.2569
28	.7171	.6030	.5403	.4992	.4696	.4471	.4146	.3752	.3248	.2516
29	.7155	.6011	.5382	.4969	.4671	.4444	.4117	.3720	.3211	.2466
30	.7141	.5994	.5362	.4947	.4648	.4420	.4090	.3691	.3176	.2419
60	.6933	.5738	.5073	.4632	.4311	.4064	.3702	.3255	.2654	.1644
∞	.6729	.5486	.4787	4319	3974	3706	3309	2804	.2085	0

[1] From Table VI, R. A. Fisher, *Statistical Methods for Research Workers*. This table is printed here through the courtesy of Dr. Fisher and his publishers, Oliver and Boyd, of Edinburgh.

Appendix Table VIII

Squares of the Natural Numbers from 100 *to* 999

n	SQUARE OF n	$n+1$	$n+2$	$n+3$	$n+4$	$n+5$	$n+6$	$n+7$	$n+8$	$n+9$
100	1 00 00	1 02 01	1 04 04	1 06 09	1 08 16	1 10 25	1 12 36	1 14 49	1 16 64	1 18 81
110	1 21 00	1 23 21	1 25 44	1 27 69	1 29 96	1 32 25	1 34 56	1 36 89	1 39 24	1 41 61
120	1 44 00	1 46 41	1 48 84	1 51 29	1 53 76	1 56 25	1 58 76	1 61 29	1 63 84	1 66 41
130	1 69 00	1 71 61	1 74 24	1 76 89	1 79 56	1 82 25	1 84 96	1 87 69	1 90 44	1 93 21
140	1 96 00	1 98 81	2 01 64	2 04 49	2 07 36	2 10 25	2 13 16	2 16 09	2 19 04	2 22 01
150	2 25 00	2 28 01	2 31 04	2 34 09	2 37 16	2 40 25	2 43 36	2 46 49	2 49 64	2 52 81
160	2 56 00	2 59 21	2 62 44	2 65 69	2 68 96	2 72 25	2 75 56	2 78 89	2 82 24	2 85 61
170	2 89 00	2 92 41	2 95 84	2 99 29	3 02 76	3 06 25	3 09 76	3 13 29	3 16 84	3 20 41
180	3 24 00	3 27 61	3 31 24	3 34 89	3 38 56	3 42 25	3 45 96	3 49 69	3 53 44	3 57 21
190	3 61 00	3 64 81	3 68 64	3 72 49	3 76 36	3 80 25	3 84 16	3 88 09	3 92 04	3 96 01
200	4 00 00	4 04 01	4 08 04	4 12 09	4 16 16	4 20 25	4 24 36	4 28 49	4 32 64	4 36 81
210	4 41 00	4 45 21	4 49 44	4 53 69	4 57 96	4 62 25	4 66 56	4 70 89	4 75 24	4 79 61
220	4 84 00	4 88 41	4 92 84	4 97 29	5 01 76	5 06 25	5 10 76	5 15 29	5 19 84	5 24 41
230	5 29 00	5 33 61	5 38 24	5 42 89	5 47 56	5 52 25	5 56 96	5 61 69	5 66 44	5 71 21
240	5 76 00	5 80 81	5 85 64	5 90 49	5 95 36	6 00 25	6 05 16	6 10 09	6 15 04	6 20 01
250	6 25 00	6 30 01	6 35 04	6 40 09	6 45 16	6 50 25	6 55 36	6 60 49	6 65 64	6 70 81
260	6 76 00	6 81 21	6 86 44	6 91 69	6 96 96	7 02 25	7 07 56	7 12 89	7 18 24	7 23 61
270	7 29 00	7 34 41	7 39 84	7 45 29	7 50 76	7 56 25	7 61 76	7 67 29	7 72 84	7 78 41
280	7 84 00	7 89 61	7 95 24	8 00 89	8 06 56	8 12 25	8 17 96	8 23 69	8 29 44	8 35 21
290	8 41 00	8 46 81	8 52 64	8 58 49	8 64 36	8 70 25	8 76 16	8 82 09	8 88 04	8 94 01
300	9 00 00	9 06 01	9 12 04	9 18 09	9 24 16	9 30 25	9 36 36	9 42 49	9 48 64	9 54 81
310	9 61 00	9 67 21	9 73 44	9 79 69	9 85 96	9 92 25	9 98 56	10 04 89	10 11 24	10 17 61
320	10 24 00	10 30 41	10 36 84	10 43 29	10 49 76	10 56 25	10 62 76	10 69 29	10 75 84	10 82 41
330	10 89 00	10 95 61	11 02 24	11 08 89	11 15 56	11 22 25	11 28 96	11 35 69	11 42 44	11 49 21
340	11 56 00	11 62 81	11 69 64	11 76 49	11 83 36	11 90 25	11 97 16	12 04 09	12 11 04	12 18 01
350	12 25 00	12 32 01	12 39 04	12 46 09	12 53 16	12 60 25	12 67 36	12 74 49	12 81 64	12 88 81
360	12 96 00	13 03 21	13 10 44	13 17 69	13 24 96	13 32 25	13 39 56	13 46 89	13 54 24	13 61 61
370	13 69 00	13 76 41	13 83 84	13 91 29	13 98 76	14 06 25	14 13 76	14 21 29	14 28 84	14 36 41
380	14 44 00	14 51 61	14 59 24	14 66 89	14 74 56	14 82 25	14 89 96	14 97 69	15 05 44	15 13 21
390	15 21 00	15 28 81	15 36 64	15 44 49	15 52 36	15 60 25	15 68 16	15 76 09	15 84 04	15 92 01
400	16 00 00	16 08 01	16 16 04	16 24 09	16 32 16	16 40 25	16 48 36	16 56 49	16 64 64	16 72 81
410	16 81 00	16 89 21	16 97 44	17 05 69	17 13 96	17 22 25	17 30 56	17 38 89	17 47 24	17 55 61
420	17 64 00	17 72 41	17 80 84	17 89 29	17 97 76	18 06 25	18 14 76	18 23 29	18 31 84	18 40 41
430	18 49 00	18 57 61	18 66 24	18 74 89	18 83 56	18 92 25	19 00 96	19 09 69	19 18 44	19 27 21
440	19 36 00	19 44 81	19 53 64	19 62 49	19 71 36	19 80 25	19 89 16	19 98 09	20 07 04	20 16 01
450	20 25 00	20 34 01	20 43 04	20 52 09	20 61 16	20 70 25	20 79 36	20 88 49	20 97 64	21 06 81
460	21 16 00	21 25 21	21 34 44	21 43 69	21 52 96	21 62 25	21 71 56	21 80 89	21 90 24	21 99 61
470	22 09 00	22 18 41	22 27 84	22 37 29	22 46 76	22 56 25	22 65 76	22 75 29	22 84 84	22 94 41
480	23 04 00	23 13 61	23 23 24	23 32 89	23 42 56	23 52 25	23 61 96	23 71 69	23 81 44	23 91 21
490	24 01 00	24 10 81	24 20 64	24 30 49	24 40 36	24 50 25	24 60 16	24 70 09	24 80 04	24 90 01
500	25 00 00	25 10 01	25 20 04	25 30 09	25 40 16	25 50 25	25 60 36	25 70 49	25 80 64	25 90 81
510	26 01 00	26 11 21	26 21 44	26 31 69	26 41 96	26 52 25	26 62 56	26 72 89	26 83 24	26 93 61
520	27 04 00	27 14 41	27 24 84	27 35 29	27 45 76	27 56 25	27 66 76	27 77 29	27 87 84	27 98 41
530	28 09 00	28 19 61	28 30 24	28 40 89	28 51 56	28 62 25	28 72 96	28 83 69	28 94 44	29 05 21
540	29 16 00	29 26 81	29 37 64	29 48 49	29 59 36	29 70 25	29 81 16	29 92 09	30 03 04	30 14 01

APPENDIX TABLE VIII—*Continued*

Squares of the Natural Numbers from 100 *to* 999

	← SQUARE OF →								
n	n	$n+1$	$n+2$	$n+3$	$n+4$	$n+5$	$n+6$	$n+7$	$n+8$	$n+9$
550	30 25 00	30 36 01	30 47 04	30 58 09	30 69 16	30 80 25	30 91 36	31 02 49	31 13 64	31 24 81
560	31 36 00	31 47 21	31 58 44	31 69 69	31 80 96	31 92 25	32 03 56	32 14 89	32 26 24	32 37 61
570	32 49 00	32 60 41	32 71 84	32 83 29	32 94 76	33 06 25	33 17 76	33 29 29	33 40 84	33 52 41
580	33 64 00	33 75 61	33 87 24	33 98 89	34 10 56	34 22 25	34 33 96	34 45 69	34 57 44	34 69 21
590	34 81 00	34 92 81	35 04 64	35 16 49	35 28 36	35 40 25	35 52 16	35 64 09	35 76 04	35 88 01
600	36 00 00	36 12 01	36 24 04	36 36 09	36 48 16	36 60 25	36 72 36	36 84 49	36 96 64	37 08 81
610	37 21 00	37 33 21	37 45 44	37 57 69	37 69 96	37 82 25	37 94 56	38 06 89	38 19 24	38 31 61
620	38 44 00	38 56 41	38 68 84	38 81 29	38 93 76	39 06 25	39 18 76	39 31 29	39 43 84	39 56 41
630	39 69 00	39 81 61	39 94 24	40 06 89	40 19 56	40 32 25	40 44 96	40 57 69	40 70 44	40 83 21
640	40 96 00	41 08 81	41 21 64	41 34 49	41 47 36	41 60 25	41 73 16	41 86 09	41 99 04	42 12 01
650	42 25 00	42 38 01	42 51 04	42 64 09	42 77 16	42 90 25	43 03 36	43 16 49	43 29 64	43 42 81
660	43 56 00	43 69 21	43 82 44	43 95 69	44 08 96	44 22 25	44 35 56	44 48 89	44 62 24	44 75 61
670	44 89 00	45 02 41	45 15 84	45 29 29	45 42 76	45 56 25	45 69 76	45 83 29	45 96 84	46 10 41
680	46 24 00	46 37 61	46 51 24	46 64 89	46 78 56	46 92 25	47 05 96	47 19 69	47 33 44	47 47 21
690	47 61 00	47 74 81	47 88 64	48 02 49	48 16 36	48 30 25	48 44 16	48 58 09	48 72 04	48 86 01
700	49 00 00	49 14 01	49 28 04	49 42 09	49 56 16	49 70 25	49 84 36	49 98 49	50 12 64	50 26 81
710	50 41 00	50 55 21	50 69 44	50 83 69	50 97 96	51 12 25	51 26 56	51 40 89	51 55 24	51 69 61
720	51 84 00	51 98 41	52 12 84	52 27 29	52 41 76	52 56 25	52 70 76	52 85 29	52 99 84	53 14 41
730	53 29 00	53 43 61	53 58 24	53 72 89	53 87 56	54 02 25	54 16 96	54 31 69	54 46 44	54 61 21
740	54 76 00	54 90 81	55 05 64	55 20 49	55 35 36	55 50 25	55 65 16	55 80 09	55 95 04	56 10 01
750	56 25 00	56 40 01	56 55 04	56 70 09	56 85 16	57 00 25	57 15 36	57 30 49	57 45 64	57 60 81
760	57 76 00	57 91 21	58 06 44	58 21 69	58 36 96	58 52 25	58 67 56	58 82 89	58 98 24	59 13 61
770	59 29 00	59 44 41	59 59 84	59 75 29	59 90 76	60 06 25	60 21 76	60 37 29	60 52 84	60 68 41
780	60 84 00	60 99 61	61 15 24	61 30 89	61 46 56	61 62 25	61 77 96	61 93 69	62 09 44	62 25 21
790	62 41 00	62 56 81	62 72 64	62 88 49	63 04 36	63 20 25	63 36 16	63 52 09	63 68 04	63 84 01
800	64 00 00	64 16 01	64 32 04	64 48 09	64 64 16	64 80 25	64 96 36	65 12 49	65 28 64	65 44 81
810	65 61 00	65 77 21	65 93 44	66 09 69	66 25 96	66 42 25	66 58 56	66 74 89	66 91 24	67 07 61
820	67 24 00	67 40 41	67 56 84	67 73 29	67 89 76	68 06 25	68 22 76	68 39 29	68 55 84	68 72 41
830	68 89 00	69 05 61	69 22 24	69 38 89	69 55 56	69 72 25	69 88 96	70 05 69	70 22 44	70 39 21
840	70 56 00	70 72 81	70 89 64	71 06 49	71 23 36	71 40 25	71 57 16	71 74 09	71 91 04	72 08 01
850	72 25 00	72 42 01	72 59 04	72 76 09	72 93 16	73 10 25	73 27 36	73 44 49	73 61 64	73 78 81
860	73 96 00	74 13 21	74 30 44	74 47 69	74 64 96	74 82 25	74 99 56	75 16 89	75 34 24	75 51 61
870	75 69 00	75 86 41	76 03 84	76 21 29	76 38 76	76 56 25	76 73 76	76 91 29	77 08 84	77 26 41
880	77 44 00	77 61 61	77 79 24	77 96 89	78 14 56	78 32 25	78 49 96	78 67 69	78 85 44	79 03 21
890	79 21 00	79 38 81	79 56 64	79 74 49	79 92 36	80 10 25	80 28 16	80 46 09	80 64 04	80 82 01
900	81 00 00	81 18 01	81 36 04	81 54 09	81 72 16	81 90 25	82 08 36	82 26 49	82 44 64	82 62 81
910	82 81 00	82 99 21	83 17 44	83 35 69	83 53 96	83 72 25	83 90 56	84 08 89	84 27 24	84 45 61
920	84 64 00	84 82 41	85 00 84	85 19 29	85 37 76	85 56 25	85 74 76	85 93 29	86 11 84	86 30 41
930	86 49 00	86 67 61	86 86 24	87 04 89	87 23 56	87 42 25	87 60 96	87 79 69	87 98 44	88 17 21
940	88 36 00	88 54 81	88 73 64	88 92 49	89 11 36	89 30 25	89 49 16	89 68 09	89 87 04	90 06 01
950	90 25 00	90 44 01	90 63 04	90 82 09	91 01 16	91 20 25	91 39 36	91 58 49	91 77 64	91 96 81
960	92 16 00	92 35 21	92 54 44	92 73 69	92 92 96	93 12 25	93 31 56	93 50 89	93 70 24	93 89 61
970	94 09 00	94 28 41	94 47 84	94 67 29	94 86 76	95 06 25	95 25 76	95 45 29	95 64 84	95 84 41
980	96 04 00	96 23 61	96 43 24	96 62 89	96 82 56	97 02 25	97 21 96	97 41 69	97 61 44	97 81 21
990	98 01 00	98 20 81	98 40 64	98 60 49	98 80 36	99 00 25	99 20 16	99 40 09	99 60 04	99 80 01

Appendix Table IX

Sums of the First Three Powers of the Natural Numbers from 1 to 50

n	$\Sigma(n)$	$\Sigma(n^2)$	$\Sigma(n^3)$	n	$\Sigma(n)$	$\Sigma(n^2)$	$\Sigma(n^3)$
1	1	1	1	26	351	6 201	123 201
2	3	5	9	27	378	6 930	142 884
3	6	14	36	28	406	7 714	164 836
4	10	30	100	29	435	8 555	189 225
5	15	55	225	30	465	9 455	216 225
6	21	91	441	31	496	10 416	246 016
7	28	140	784	32	528	11 440	278 784
8	36	204	1 296	33	561	12 529	314 721
9	45	285	2 025	34	595	13 685	354 025
10	55	385	3 025	35	630	14 910	396 900
11	66	506	4 356	36	666	16 206	443 556
12	78	650	6 084	37	703	17 575	494 209
13	91	819	8 281	38	741	19 019	549 081
14	105	1 015	11 025	39	780	20 540	608 400
15	120	1 240	14 400	40	820	22 140	672 400
16	136	1 496	18 496	41	861	23 821	741 321
17	153	1 785	23 409	42	903	25 585	815 409
18	171	2 109	29 241	43	946	27 434	894 916
19	190	2 470	36 100	44	990	29 370	980 100
20	210	2 870	44 100	45	1 035	31 395	1 071 225
21	231	3 311	53 361	46	1 081	33 511	1 168 561
22	253	3 795	64 009	47	1 128	35 720	1 272 384
23	276	4 324	76 176	48	1 176	38 024	1 382 976
24	300	4 900	90 000	49	1 225	40 425	1 500 625
25	325	5 525	105 625	50	1 275	42 925	1 625 625

Appendix Table X

Five-Place Logarithms of Numbers

100–150

N	L	0	1	2	3	4	5	6	7	8	9
100		00 000	043	087	130	173	217	260	303	346	389
101		432	475	518	561	604	647	689	732	775	817
102		860	903	945	988	*030	*072	*115	*157	*199	*242
103		01 284	326	368	410	452	494	536	578	620	662
104		703	745	787	828	870	912	953	995	*036	*078
105		02 119	160	202	243	284	325	366	407	449	490
106		531	572	612	653	694	735	776	816	857	898
107		938	979	*019	*060	*100	*141	*181	*222	*262	*302
108		03 342	383	423	463	503	543	583	623	663	703
109		743	782	822	862	902	941	981	*021	*060	*100
110		04 139	179	218	258	297	336	376	415	454	493
111		532	571	610	650	689	727	766	805	844	883
112		922	961	999	*038	*077	*115	*154	*192	*231	*269
113		05 308	346	385	423	461	500	538	576	614	652
114		690	729	767	805	843	881	918	956	994	*032
115		06 070	108	145	183	221	258	296	333	371	408
116		446	483	521	558	595	633	670	707	744	781
117		819	856	893	930	967	*004	*041	*078	*115	*151
118		07 188	225	262	298	335	372	408	445	482	518
119		555	591	628	664	700	737	773	809	846	882
120		918	954	990	*027	*063	*099	*135	*171	*207	*243
121		08 279	314	350	386	422	458	493	529	565	600
122		636	672	707	743	778	814	849	884	920	955
123		991	*026	*061	*096	*132	*167	*202	*237	*272	*307
124		09 342	377	412	447	482	517	552	587	621	656
125		691	726	760	795	830	864	899	934	968	*003
126		10 037	072	106	140	175	209	243	278	312	346
127		380	415	449	483	517	551	585	619	653	687
128		721	755	789	823	857	890	924	958	992	*025
129		11 059	093	126	160	193	227	261	294	327	361
130		394	428	461	494	528	561	594	628	661	694
131		727	760	793	826	860	893	926	959	992	*024
132		12 057	090	123	156	189	222	254	287	320	352
133		385	418	450	483	516	548	581	613	646	678
134		710	743	775	808	840	872	905	937	969	*001
135		13 033	066	098	130	162	194	226	258	290	322
136		354	386	418	450	481	513	545	577	609	640
137		672	704	735	767	799	830	862	893	925	956
138		988	*019	*051	*082	*114	*145	*176	*208	*239	*270
139		14 301	333	364	395	426	457	489	520	551	582
140		613	644	675	706	737	768	799	829	860	891
141		922	953	983	*014	*045	*076	*106	*137	*168	*198
142		15 229	259	290	320	351	381	412	442	473	503
143		534	564	594	625	655	685	715	746	776	806
144		836	866	897	927	957	987	*017	*047	*077	*107
145		16 137	167	197	227	256	286	316	346	376	406
146		435	465	495	524	554	584	613	643	673	702
147		732	761	791	820	850	879	909	938	967	997
148		17 026	056	085	114	143	173	202	231	260	289
149		319	348	377	406	435	464	493	522	551	580
150		609	638	667	696	725	754	782	811	840	869
N	L	0	1	2	3	4	5	6	7	8	9

Prop. Pts.			
	44	43	42
1	4.4	4.3	4.2
2	8.8	8.6	8.4
3	13.2	12.9	12.6
4	17.6	17.2	16.8
5	22.0	21.5	21.0
6	26.4	25.8	25.2
7	30.8	30.1	29.4
8	35.2	34.4	33.6
9	39.6	38.7	37.8
	41	40	39
1	4.1	4.0	3.9
2	8.2	8.0	7.8
3	12.3	12.0	11.7
4	16.4	16.0	15.6
5	20.5	20.0	19.5
6	24.6	24.0	23.4
7	28.7	28.0	27.3
8	32.8	32.0	31.2
9	36.9	36.0	35.1
	38	37	36
1	3.8	3.7	3.6
2	7.6	7.4	7.2
3	11.4	11.1	10.8
4	15.2	14.8	14.4
5	19.0	18.5	18.0
6	22.8	22.2	21.6
7	26.6	25.9	25.2
8	30.4	29.6	28.8
9	34.2	33.3	32.4
	35	34	33
1	3.5	3.4	3.3
2	7.0	6.8	6.6
3	10.5	10.2	9.9
4	14.0	13.6	13.2
5	17.5	17.0	16.5
6	21.0	20.4	19.8
7	24.5	23.8	23.1
8	28.0	27.2	26.4
9	31.5	30.6	29.7
	32	31	30
1	3.2	3.1	3.0
2	6.4	6.2	6.0
3	9.6	9.3	9.0
4	12.8	12.4	12.0
5	16.0	15.5	15.0
6	19.2	18.6	18.0
7	22.4	21.7	21.0
8	25.6	24.8	24.0
9	28.8	27.9	27.0

Appendix Table X

Five-Place Logarithms of Numbers

150–200

N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.
150	17 609	638	667	696	725	754	782	811	840	869	
151	898	926	955	984	*013	*041	*070	*099	*127	*156	
152	18 184	213	241	270	298	327	355	384	412	441	29 28
153	469	498	526	554	583	611	639	667	696	724	
154	752	780	808	837	865	893	921	949	977	*005	1 2.9 2.8
											2 5.8 5.6
155	19 033	061	089	117	145	173	201	229	257	285	3 8.7 8.4
156	312	340	368	396	424	451	479	507	535	562	4 11.6 11.2
157	590	618	645	673	700	728	756	783	811	838	5 14.5 14.0
158	866	893	921	948	976	*003	*030	*058	*085	*112	6 17.4 16.8
159	20 140	167	194	222	249	276	303	330	358	385	7 20.3 19.6
											8 23.2 22.4
160	412	439	466	493	520	548	575	602	629	656	9 26.1 25.2
161	683	710	737	763	790	817	844	871	898	925	
162	952	978	*005	*032	*059	*085	*112	*139	*165	*192	27 26
163	21 219	245	272	299	325	352	378	405	431	458	
164	484	511	537	564	590	617	643	669	696	722	1 2.7 2.6
											2 5.4 5.2
165	748	775	801	827	854	880	906	932	958	985	3 8.1 7.8
166	22 011	037	063	089	115	141	167	194	220	246	4 10.8 10.4
167	272	298	324	350	376	401	427	453	479	505	5 13.5 13.0
168	531	557	583	608	634	660	686	712	737	763	6 16.2 15.6
169	789	814	840	866	891	917	943	968	994	*019	7 18.9 18.2
											8 21.6 20.8
170	23 045	070	096	121	147	172	198	223	249	274	9 24.3 23.4
171	300	325	350	376	401	426	452	477	502	528	
172	553	578	603	629	654	679	704	729	754	779	25
173	805	830	855	880	905	930	955	980	*005	*030	1 2.5
174	24 055	080	105	130	155	180	204	229	254	279	2 5.0
											3 7.5
175	304	329	353	378	403	428	452	477	502	527	4 10.0
176	551	576	601	625	650	674	699	724	748	773	5 12.5
177	797	822	846	871	895	920	944	969	993	*018	6 15.0
178	25 042	066	091	115	139	164	188	212	237	261	7 17.5
179	285	310	334	358	382	406	431	455	479	503	8 20.0
											9 22.5
180	527	551	575	600	624	648	672	696	720	744	
181	768	792	816	840	864	888	912	935	959	983	24 23
182	26 007	031	055	079	102	126	150	174	198	221	
183	245	269	293	316	340	364	387	411	435	458	1 2.4 2.3
184	482	505	529	553	576	600	623	647	670	694	2 4.8 4.6
											3 7.2 6.9
185	717	741	764	788	811	834	858	881	905	928	4 9.6 9.2
186	951	975	998	*021	*045	*068	*091	*114	*138	*161	5 12.0 11.5
187	27 184	207	231	254	277	300	323	346	370	393	6 14.4 13.8
188	416	439	462	485	508	531	554	577	600	623	7 16.8 16.1
189	646	669	692	715	738	761	784	807	830	852	8 19.2 18.4
											9 21.6 20.7
190	875	898	921	944	967	989	*012	*035	*058	*081	
191	28 103	126	149	171	194	217	240	262	285	307	22 21
192	330	353	375	398	421	443	466	488	511	533	
193	556	578	601	623	646	668	691	713	735	758	1 2.2 2.1
194	780	803	825	847	870	892	914	937	959	981	2 4.4 4.2
											3 6.6 6.3
195	29 003	026	048	070	092	115	137	159	181	203	4 8.8 8.4
196	226	248	270	292	314	336	358	380	403	425	5 11.0 10.5
197	447	469	491	513	535	557	579	601	623	645	6 13.2 12.6
198	667	688	710	732	754	776	798	820	842	863	7 15.4 14.7
199	885	907	929	951	973	994	*016	*038	*060	*081	8 17.6 16.8
											9 19.8 18.9
200	30 103	125	146	168	190	211	233	255	276	298	
N	**L 0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Prop. Pts.**

Appendix Table X

Five-Place Logarithms of Numbers

200–250

N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.
200	30 103	125	146	168	190	211	233	255	276	298	
201	320	341	363	384	406	428	449	471	492	514	
202	535	557	578	600	621	643	664	685	707	728	22 | 21
203	750	771	792	814	835	856	878	899	920	942	
204	963	984	*006	*027	*048	*069	*091	*112	*133	*154	1 | 2.2 | 2.1
205	31 175	197	218	239	260	281	302	323	345	366	2 | 4.4 | 4.2
206	387	408	429	450	471	492	513	534	555	576	3 | 6.6 | 6.3
207	597	618	639	660	681	702	723	744	765	785	4 | 8.8 | 8.4
208	806	827	848	869	890	911	931	952	973	994	5 | 11.0 | 10.5
209	32 015	035	056	077	098	118	139	160	181	201	6 | 13.2 | 12.6
											7 | 15.4 | 14.7
											8 | 17.6 | 16.8
											9 | 19.8 | 18.9
210	222	243	263	284	305	325	346	366	387	408	
211	428	449	469	490	510	531	552	572	593	613	
212	634	654	675	695	715	736	756	777	797	818	20
213	838	858	879	899	919	940	960	980	*001	*021	
214	33 041	062	082	102	122	143	163	183	203	224	1 | 2.0
215	244	264	284	304	325	345	365	385	405	425	2 | 4.0
216	445	465	486	506	526	546	566	586	606	626	3 | 6.0
217	646	666	686	706	726	746	766	786	806	826	4 | 8.0
218	846	866	885	905	925	945	965	985	*005	*025	5 | 10.0
219	34 044	064	084	104	124	143	163	183	203	223	6 | 12.0
											7 | 14.0
											8 | 16.0
											9 | 18.0
220	242	262	282	301	321	341	361	380	400	420	
221	439	459	479	498	518	537	557	577	596	616	
222	635	655	674	694	713	733	753	772	792	811	19
223	830	850	869	889	908	928	947	967	986	*005	
224	35 025	044	064	083	102	122	141	160	180	199	1 | 1.9
225	218	238	257	276	295	315	334	353	372	392	2 | 3.8
226	411	430	449	468	488	507	526	545	564	583	3 | 5.7
227	603	622	641	660	679	698	717	736	755	774	4 | 7.6
228	793	813	832	851	870	889	908	927	946	965	5 | 9.5
229	984	*003	*021	*040	*059	*078	*097	*116	*135	*154	6 | 11.4
											7 | 13.3
											8 | 15.2
											9 | 17.1
230	36 173	192	211	229	248	267	286	305	324	342	
231	361	380	399	418	436	455	474	493	511	530	
232	549	568	586	605	624	642	661	680	698	717	18
233	736	754	773	791	810	829	847	866	884	903	
234	922	940	959	977	996	*014	*033	*051	*070	*088	1 | 1.8
235	37 107	125	144	162	181	199	218	236	254	273	2 | 3.6
236	291	310	328	346	365	383	401	420	438	457	3 | 5.4
237	475	493	511	530	548	566	585	603	621	639	4 | 7.2
238	658	676	694	712	731	749	767	785	803	822	5 | 9.0
239	840	858	876	894	912	931	949	967	985	*003	6 | 10.8
											7 | 12.6
											8 | 14.4
											9 | 16.2
240	38 021	039	057	075	093	112	130	148	166	184	
241	202	220	238	256	274	292	310	328	346	364	
242	382	399	417	435	453	471	489	507	525	543	17
243	561	578	596	614	632	650	668	686	703	721	
244	739	757	775	792	810	828	846	863	881	899	1 | 1.7
245	917	934	952	970	987	*005	*023	*041	*058	*076	2 | 3.4
246	39 094	111	129	146	164	182	199	217	235	252	3 | 5.1
247	270	287	305	322	340	358	375	393	410	428	4 | 6.8
248	445	463	480	498	515	533	550	568	585	602	5 | 8.5
249	620	637	655	672	690	707	724	742	759	777	6 | 10.2
											7 | 11.9
											8 | 13.6
											9 | 15.3
250	794	811	829	846	863	881	898	915	933	950	
N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.

Appendix Table X

Five-Place Logarithms of Numbers

250–300

N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.
250	39 794	811	829	846	863	881	898	915	933	950	
251	967	985	*002	*019	*037	*054	*071	*088	*106	*123	
252	40 140	157	175	192	209	226	243	261	278	295	18
253	312	329	346	364	381	398	415	432	449	466	
254	483	500	518	535	552	569	586	603	620	637	1 1.8
											2 3.6
255	654	671	688	705	722	739	756	773	790	807	3 5.4
256	824	841	858	875	892	909	926	943	960	976	4 7.2
257	993	*010	*027	*044	*061	*078	*095	*111	*128	*145	5 9.0
258	41 162	179	196	212	229	246	263	280	296	313	6 10.8
259	330	347	363	380	397	414	430	447	464	481	7 12.6
											8 14.4
											9 16.2
260	497	514	531	547	564	581	597	614	631	647	
261	664	681	697	714	731	747	764	780	797	814	
262	830	847	863	880	896	913	929	946	963	979	17
263	996	*012	*029	*045	*062	*078	*095	*111	*127	*144	
264	42 160	177	193	210	226	243	259	275	292	308	1 1.7
											2 3.4
265	325	341	357	374	390	406	423	439	455	472	3 5.1
266	488	504	521	537	553	570	586	602	619	635	4 6.8
267	651	667	684	700	716	732	749	765	781	797	5 8.5
268	813	830	846	862	878	894	911	927	943	959	6 10.2
269	975	991	*008	*024	*040	*056	*072	*088	*104	*120	7 11.9
											8 13.6
											9 15.3
270	43 136	152	169	185	201	217	233	249	265	281	
271	297	313	329	345	361	377	393	409	425	441	
272	457	473	489	505	521	537	553	569	584	600	16
273	616	632	648	664	680	696	712	727	743	759	
274	775	791	807	823	838	854	870	886	902	917	1 1.6
											2 3.2
275	933	949	965	981	996	*012	*028	*044	*059	*075	3 4.8
276	44 091	107	122	138	154	170	185	201	217	232	4 6.4
277	248	264	279	295	311	326	342	358	373	389	5 8.0
278	404	420	436	451	467	483	498	514	529	545	6 9.6
279	560	576	592	607	623	638	654	669	685	700	7 11.2
											8 12.8
											9 14.4
280	716	731	747	762	778	793	809	824	840	855	
281	871	886	902	917	932	948	963	979	994	*010	
282	45 025	040	056	071	086	102	117	133	148	163	15
283	179	194	209	225	240	255	271	286	301	317	
284	332	347	362	378	393	408	423	439	454	469	1 1.5
											2 3.0
285	484	500	515	530	545	561	576	591	606	621	3 4.5
286	637	652	667	682	697	712	728	743	758	773	4 6.0
287	788	803	818	834	849	864	879	894	909	924	5 7.5
288	939	954	969	984	*000	*015	*030	*045	*060	*075	6 9.0
289	46 090	105	120	135	150	165	180	195	210	225	7 10.5
											8 12.0
											9 13.5
290	240	255	270	285	300	315	330	345	359	374	
291	389	404	419	434	449	464	479	494	509	523	
292	538	553	568	583	598	613	627	642	657	672	14
293	687	702	716	731	746	761	776	790	805	820	
294	835	850	864	879	894	909	923	938	953	967	1 1.4
											2 2.8
295	982	997	*012	*026	*041	*056	*070	*085	*100	*114	3 4.2
296	47 129	144	159	173	188	202	217	232	246	261	4 5.6
297	276	290	305	319	334	349	363	378	392	407	5 7.0
298	422	436	451	465	480	494	509	524	538	553	6 8.4
299	567	582	596	611	625	640	654	669	683	698	7 9.8
											8 11.2
											9 12.6
300	712	727	741	756	770	784	799	813	828	842	
N	**L 0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Prop. Pts.**

$\log e = .43429$

Appendix Table X

Five-Place Logarithms of Numbers

300–350

N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.
300	47 712	727	741	756	770	784	799	813	828	842	
301	857	871	885	900	914	929	943	958	972	986	
302	48 001	015	029	044	058	073	087	101	116	130	
303	144	159	173	187	202	216	230	244	259	273	
304	287	302	316	330	344	359	373	387	401	416	**15**
305	430	444	458	473	487	501	515	530	544	558	1 1.5
306	572	586	601	615	629	643	657	671	686	700	2 3.0
307	714	728	742	756	770	785	799	813	827	841	3 4.5
308	855	869	883	897	911	926	940	954	968	982	4 6.0
309	996	*010	*024	*038	*052	*066	*080	*094	*108	*122	5 7.5
											6 9.0
											7 10.5
310	49 136	150	164	178	192	206	220	234	248	262	8 12.0
											9 13.5
311	276	290	304	318	332	346	360	374	388	402	
312	415	429	443	457	471	485	499	513	527	541	
313	554	568	582	596	610	624	638	651	665	679	
314	693	707	721	734	748	762	776	790	803	817	
315	831	845	859	872	886	900	914	927	941	955	
316	969	982	996	*010	*024	*037	*051	*065	*079	*092	**14**
317	50 106	120	133	147	161	174	188	202	215	229	1 1.4
318	243	256	270	284	297	311	325	338	352	365	2 2.8
319	379	393	406	420	433	447	461	474	488	501	3 4.2
											4 5.6
320	515	529	542	556	569	583	596	610	623	637	5 7.0
											6 8.4
321	651	664	678	691	705	718	732	745	759	772	7 9.8
322	786	799	813	826	840	853	866	880	893	907	8 11.2
323	920	934	947	961	974	987	*001	*014	*028	*041	9 12.6
324	51 055	068	081	095	108	121	135	148	162	175	
325	188	202	215	228	242	255	268	282	295	308	
326	322	335	348	362	375	388	402	415	428	441	
327	455	468	481	495	508	521	534	548	561	574	
328	587	601	614	627	640	654	667	680	693	706	**13**
329	720	733	746	759	772	786	799	812	825	838	1 1.3
330	851	865	878	891	904	917	930	943	957	970	2 2.6
											3 3.9
331	983	996	*009	*022	*035	*048	*061	*075	*088	*101	4 5.2
332	52 114	127	140	153	166	179	192	205	218	231	5 6.5
333	244	257	270	284	297	310	323	336	349	362	6 7.8
334	375	388	401	414	427	440	453	466	479	492	7 9.1
											8 10.4
											9 11.7
335	504	517	530	543	556	569	582	595	608	621	
336	634	647	660	673	686	699	711	724	737	750	
337	763	776	789	802	815	827	840	853	866	879	
338	892	905	917	930	943	956	969	982	994	*007	
339	53 020	033	046	058	071	084	097	110	122	135	
											12
340	148	161	173	186	199	212	224	237	250	263	
											1 1.2
341	275	288	301	314	326	339	352	364	377	390	2 2.4
342	403	415	428	441	453	466	479	491	504	517	3 3.6
343	529	542	555	567	580	593	605	618	631	643	4 4.8
344	656	668	681	694	706	719	732	744	757	769	5 6.0
											6 7.2
											7 8.4
345	782	794	807	820	832	845	857	870	882	895	8 9.6
346	908	920	933	945	958	970	983	995	*008	*020	9 10.8
347	54 033	045	058	070	083	095	108	120	133	145	
348	158	170	183	195	208	220	233	245	258	270	
349	283	295	307	320	332	345	357	370	382	394	
350	407	419	432	444	456	469	481	494	506	518	
N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.

$\log \pi = .49715$

Appendix Table X

Five-Place Logarithms of Numbers

350–400

N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.
350	54 407	419	432	444	456	469	481	494	506	518	
351	531	543	555	568	580	593	605	617	630	642	
352	654	667	679	691	704	716	728	741	753	765	
353	777	790	802	814	827	839	851	864	876	888	
354	900	913	925	937	949	962	974	986	998	*011	13
355	55 023	035	047	060	072	084	096	108	121	133	1 1.3
356	145	157	169	182	194	206	218	230	242	255	2 2.6
357	267	279	291	303	315	328	340	352	364	376	3 3.9
358	388	400	413	425	437	449	461	473	485	497	4 5.2
359	509	522	534	546	558	570	582	594	606	618	5 6.5
360	630	642	654	666	678	691	703	715	727	739	6 7.8 7 9.1 8 10.4 9 11.7
361	751	763	775	787	799	811	823	835	847	859	
362	871	883	895	907	919	931	943	955	967	979	
363	991	*003	*015	*027	*038	*050	*062	*074	*086	*098	
364	56 110	122	134	146	158	170	182	194	205	217	
365	229	241	253	265	277	289	301	312	324	336	12
366	348	360	372	384	396	407	419	431	443	455	
367	467	478	490	502	514	526	538	549	561	573	1 1.2
368	585	597	608	620	632	644	656	667	679	691	2 2.4
369	703	714	726	738	750	761	773	785	797	808	3 3.6 4 4.8 5 6.0
370	820	832	844	855	867	879	891	902	914	926	6 7.2 7 8.4
371	937	949	961	972	984	996	*008	*019	*031	*043	8 9.6
372	57 054	066	078	089	101	113	124	136	148	159	9 10.8
373	171	183	194	206	217	229	241	252	264	276	
374	287	299	310	322	334	345	357	368	380	392	
375	403	415	426	438	449	461	473	484	496	507	
376	519	530	542	553	565	576	588	600	611	623	
377	634	646	657	669	680	692	703	715	726	738	11
378	749	761	772	784	795	807	818	830	841	852	
379	864	875	887	898	910	921	933	944	955	967	1 1.1
380	978	990	*001	*013	*024	*035	*047	*058	*070	*081	2 2.2 3 3.3 4 4.4 5 5.5 6 6.6 7 7.7 8 8.8 9 9.9
381	58 092	104	115	127	138	149	161	172	184	195	
382	206	218	229	240	252	263	274	286	297	309	
383	320	331	343	354	365	377	388	399	410	422	
384	433	444	456	467	478	490	501	512	524	535	
385	546	557	569	580	591	602	614	625	636	647	
386	659	670	681	692	704	715	726	737	749	760	
387	771	782	794	805	816	827	838	850	861	872	
388	883	894	906	917	928	939	950	961	973	984	
389	995	*006	*017	*028	*040	*051	*062	*073	*084	*095	10
390	59 106	118	129	140	151	162	173	184	195	207	1 1.0
391	218	229	240	251	262	273	284	295	306	318	2 2.0
392	329	340	351	362	373	384	395	406	417	428	3 3.0
393	439	450	461	472	483	494	506	517	528	539	4 4.0
394	550	561	572	583	594	605	616	627	638	649	5 5.0 6 6.0 7 7.0
395	660	671	682	693	704	715	726	737	748	759	8 8.0
396	770	780	791	802	813	824	835	846	857	868	9 9.0
397	879	890	901	912	923	934	945	956	966	977	
398	988	999	*010	*021	*032	*043	*054	*065	*076	*086	
399	60 097	108	119	130	141	152	163	173	184	195	
400	206	217	228	239	249	260	271	282	293	304	
N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.

Appendix Table X

Five-Place Logarithms of Numbers

400–450

N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.
400	60 206	217	228	239	249	260	271	282	293	304	
401	314	325	336	347	358	369	379	390	401	412	
402	423	433	444	455	466	477	487	498	509	520	
403	531	541	552	563	574	584	595	606	617	627	
404	638	649	660	670	681	692	703	713	724	735	
405	746	756	767	778	788	799	810	821	831	842	
406	853	863	874	885	895	906	917	927	938	949	11
407	959	970	981	991	*002	*013	*023	*034	*045	*055	
408	61 066	077	087	098	109	119	130	140	151	162	1 1.1
409	172	183	194	204	215	225	236	247	257	268	2 2.2
											3 3.3
410	278	289	300	310	321	331	342	352	363	374	4 4.4
411	384	395	405	416	426	437	448	458	469	479	5 5.5
412	490	500	511	521	532	542	553	563	574	584	6 6.6
413	595	606	616	627	637	648	658	669	679	690	7 7.7
414	700	711	721	731	742	752	763	773	784	794	8 8.8
											9 9.9
415	805	815	826	836	847	857	868	878	888	899	
416	909	920	930	941	951	962	972	982	993	*003	
417	62 014	024	034	045	055	066	076	086	097	107	
418	118	128	138	149	159	170	180	190	201	211	
419	221	232	242	252	263	273	284	294	304	315	
420	325	335	346	356	366	377	387	397	408	418	
421	428	439	449	459	469	480	490	500	511	521	
422	531	542	552	562	572	583	593	603	613	624	10
423	634	644	655	665	675	685	696	706	716	726	1 1.0
424	737	747	757	767	778	788	798	808	818	829	2 2.0
											3 3.0
425	839	849	859	870	880	890	900	910	921	931	4 4.0
426	941	951	961	972	982	992	*002	*012	*022	*033	5 5.0
427	63 043	053	063	073	083	094	104	114	124	134	6 6.0
428	144	155	165	175	185	195	205	215	225	236	7 7.0
429	246	256	266	276	286	296	306	317	327	337	8 8.0
											9 9.0
430	347	357	367	377	387	397	407	417	428	438	
431	448	458	468	478	488	498	508	518	528	538	
432	548	558	568	579	589	599	609	619	629	639	
433	649	659	669	679	689	699	709	719	729	739	
434	749	759	769	779	789	799	809	819	829	839	
435	849	859	869	879	889	899	909	919	929	939	
436	949	959	969	979	988	998	*008	*018	*028	*038	
437	64 048	058	068	078	088	098	108	118	128	137	9
438	147	157	167	177	187	197	207	217	227	237	1 0.9
439	246	256	266	276	286	296	306	316	326	335	2 1.8
											3 2.7
440	345	355	365	375	385	395	404	414	424	434	4 3.6
441	444	454	464	473	483	493	503	513	523	532	5 4.5
442	542	552	562	572	582	591	601	611	621	631	6 5.4
443	640	650	660	670	680	689	699	709	719	729	7 6.3
444	738	748	758	768	777	787	797	807	816	826	8 7.2
											9 8.1
445	836	846	856	865	875	885	895	904	914	924	
446	933	943	953	963	972	982	992	*002	*011	*021	
447	65 031	040	050	060	070	079	089	099	108	118	
448	128	137	147	157	167	176	186	196	205	215	
449	225	234	244	254	263	273	283	292	302	312	
450	321	331	341	350	360	369	379	389	398	408	
N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.

Appendix Table X

Five-Place Logarithms of Numbers

450–500

N	L 0	1	2	3	4	5	6	7	8	9
450	65 321	331	341	350	360	369	379	389	398	408
451	418	427	437	447	456	466	475	485	495	504
452	514	523	533	543	552	562	571	581	591	600
453	610	619	629	639	648	658	667	677	686	696
454	706	715	725	734	744	753	763	772	782	792
455	801	811	820	830	839	849	858	868	877	887
456	896	906	916	925	935	944	954	963	973	982
457	992	*001	*011	*020	*030	*039	*049	*058	*068	*077
458	66 087	096	106	115	124	134	143	153	162	172
459	181	191	200	210	219	229	238	247	257	266
460	276	285	295	304	314	323	332	342	351	361
461	370	380	389	398	408	417	427	436	445	455
462	464	474	483	492	502	511	521	530	539	549
463	558	567	577	586	596	605	614	624	633	642
464	652	661	671	680	689	699	708	717	727	736
465	745	755	764	773	783	792	801	811	820	829
466	839	848	857	867	876	885	894	904	913	922
467	932	941	950	960	969	978	987	997	*006	*015
468	67 025	034	043	052	062	071	080	089	099	108
469	117	127	136	145	154	164	173	182	191	201
470	210	219	228	237	247	256	265	274	284	293
471	302	311	321	330	339	348	357	367	376	385
472	394	403	413	422	431	440	449	459	468	477
473	486	495	504	514	523	532	541	550	560	569
474	578	587	596	605	614	624	633	642	651	660
475	669	679	688	697	706	715	724	733	742	752
476	761	770	779	788	797	806	815	825	834	843
477	852	861	870	879	888	897	906	916	925	934
478	943	952	961	970	979	988	997	*006	*015	*024
479	68 034	043	052	061	070	079	088	097	106	115
480	124	133	142	151	160	169	178	187	196	205
481	215	224	233	242	251	260	269	278	287	296
482	305	314	323	332	341	350	359	368	377	386
483	395	404	413	422	431	440	449	458	467	476
484	485	494	502	511	520	529	538	547	556	565
485	574	583	592	601	610	619	628	637	646	655
486	664	673	681	690	699	708	717	726	735	744
487	753	762	771	780	789	797	806	815	824	833
488	842	851	860	869	878	886	895	904	913	922
489	931	940	949	958	966	975	984	993	*002	*011
490	69 020	028	037	046	055	064	073	082	090	099
491	108	117	126	135	144	152	161	170	179	188
492	197	205	214	223	232	241	249	258	267	276
493	285	294	302	311	320	329	338	346	355	364
494	373	381	390	399	408	417	425	434	443	452
495	461	469	478	487	496	504	513	522	531	539
496	548	557	566	574	583	592	601	609	618	627
497	636	644	653	662	671	679	688	697	705	714
498	723	732	740	749	758	767	775	784	793	801
499	810	819	827	836	845	854	862	871	880	888
500	897	906	914	923	932	940	949	958	906	975
N	L 0	1	2	3	4	5	6	7	8	9

Prop. Pts.	
	10
1	1.0
2	2.0
3	3.0
4	4.0
5	5.0
6	6.0
7	7.0
8	8.0
9	9.0
	9
1	0.9
2	1.8
3	2.7
4	3.6
5	4.5
6	5.4
7	6.3
8	7.2
9	8.1
	8
1	0.8
2	1.6
3	2.4
4	3.2
5	4.0
6	4.8
7	5.6
8	6.4
9	7.2

Prop. Pts.

Appendix Table X

Five-Place Logarithms of Numbers

500–550

N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.
500	69 897	906	914	923	932	940	949	958	966	975	
501	984	992	*001	*010	*018	*027	*036	*044	*053	*062	
502	70 070	079	088	096	105	114	122	131	140	148	
503	157	165	174	183	191	200	209	217	226	234	
504	243	252	260	269	278	286	295	303	312	321	
505	329	338	346	355	364	372	381	389	398	406	
506	415	424	432	441	449	458	467	475	484	492	
507	501	509	518	526	535	544	552	561	569	578	9
508	586	595	603	612	621	629	638	646	655	663	1 0.9
509	672	680	689	697	706	714	723	731	740	749	2 1.8
510	757	766	774	783	791	800	808	817	825	834	3 2.7 4 3.6 5 4.5
511	842	851	859	868	876	885	893	902	910	919	6 5.4
512	927	935	944	952	961	969	978	986	995	*003	7 6.3
513	71 012	020	029	037	046	054	063	071	079	088	8 7.2
514	096	105	113	122	130	139	147	155	164	172	9 8.1
515	181	189	198	206	214	223	231	240	248	257	
516	265	273	282	290	299	307	315	324	332	341	
517	349	357	366	374	383	391	399	408	416	425	
518	433	441	450	458	466	475	483	492	500	508	
519	517	525	533	542	550	559	567	575	584	592	
520	600	609	617	625	634	642	650	659	667	675	
521	684	692	700	709	717	725	734	742	750	759	
522	767	775	784	792	800	809	817	825	834	842	8
523	850	858	867	875	883	892	900	908	917	925	1 0.8
524	933	941	950	958	966	975	983	991	999	*008	2 1.6
525	72 016	024	032	041	049	057	066	074	082	090	3 2.4 4 3.2
526	099	107	115	123	132	140	148	156	165	173	5 4.0 6 4.8
527	181	189	198	206	214	222	230	239	247	255	7 5.6
528	263	272	280	288	296	304	313	321	329	337	8 6.4
529	346	354	362	370	378	387	395	403	411	419	9 7.2
530	428	436	444	452	460	469	477	485	493	501	
531	509	518	526	534	542	550	558	567	575	583	
532	591	599	607	616	624	632	640	648	656	665	
533	673	681	689	697	705	713	722	730	738	746	
534	754	762	770	779	787	795	803	811	819	827	
535	835	843	852	860	868	876	884	892	900	908	
536	916	925	933	941	949	957	965	973	981	989	
537	997	*006	*014	*022	*030	*038	*046	*054	*062	*070	7
538	73 078	086	094	102	111	119	127	135	143	151	1 0.7
539	159	167	175	183	191	199	207	215	223	231	2 1.4 3 2.1
540	239	247	255	263	272	280	288	296	304	312	4 2.8 5 3.5
541	320	328	336	344	352	360	368	376	384	392	6 4.2
542	400	408	416	424	432	440	448	456	464	472	7 4.9
543	480	488	496	504	512	520	528	536	544	552	8 5.6
544	560	568	576	584	592	600	608	616	624	632	9 6.3
545	640	648	656	664	672	679	687	695	703	711	
546	719	727	735	743	751	759	767	775	783	791	
547	799	807	815	823	830	838	846	854	862	870	
548	878	886	894	902	910	918	926	933	941	949	
549	957	965	973	981	989	997	*005	*013	*020	*028	
550	74 036	044	052	060	068	076	084	092	099	107	
N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.

Appendix Table X

Five-Place Logarithms of Numbers

550–600

N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.
550	74 036	044	052	060	068	076	084	092	099	107	
551	115	123	131	139	147	155	162	170	178	186	
552	194	202	210	218	225	233	241	249	257	265	
553	273	280	288	296	304	312	320	327	335	343	
554	351	359	367	374	382	390	398	406	414	421	
555	429	437	445	453	461	468	476	484	492	500	
556	507	515	523	531	539	547	554	562	570	578	
557	586	593	601	609	617	624	632	640	648	656	
558	663	671	679	687	695	702	710	718	726	733	
559	741	749	757	764	772	780	788	796	803	811	
560	819	827	834	842	850	858	865	873	881	889	
561	896	904	912	920	927	935	943	950	958	966	8
562	974	981	989	997	*005	*012	*020	*028	*035	*043	
563	75 051	059	066	074	082	089	097	105	113	120	1 0.8
564	128	136	143	151	159	166	174	182	189	197	2 1.6
565	205	213	220	228	236	243	251	259	266	274	3 2.4 4 3.2
566	282	289	297	305	312	320	328	335	343	351	5 4.0
567	358	366	374	381	389	397	404	412	420	427	6 4.8
568	435	442	450	458	465	473	481	488	496	504	7 5.6
569	511	519	526	534	542	549	557	565	572	580	8 6.4 9 7.2
570	587	595	603	610	618	626	633	641	648	656	
571	664	671	679	686	694	702	709	717	724	732	
572	740	747	755	762	770	778	785	793	800	808	
573	815	823	831	838	846	853	861	868	876	884	
574	891	899	906	914	921	929	937	944	952	959	
575	967	974	982	989	997	*005	*012	*020	*027	*035	
576	76 042	050	057	065	072	080	087	095	103	110	
577	118	125	133	140	148	155	163	170	178	185	
578	193	200	208	215	223	230	238	245	253	260	
579	268	275	283	290	298	305	313	320	328	335	
580	343	350	358	365	373	380	388	395	403	410	
581	418	425	433	440	448	455	462	470	477	485	
582	492	500	507	515	522	530	537	545	552	559	7
583	567	574	582	589	597	604	612	619	626	634	1 0.7
584	641	649	656	664	671	678	686	693	701	708	2 1.4 3 2.1
585	716	723	730	738	745	753	760	768	775	782	4 2.8
586	790	797	805	812	819	827	834	842	849	856	5 3.5
587	864	871	879	886	893	901	908	916	923	930	6 4.2
588	938	945	953	960	967	975	982	989	997	*004	7 4.9 8 5.6
589	77 012	019	026	034	041	048	056	063	070	078	9 6.3
590	085	093	100	107	115	122	129	137	144	151	
591	159	166	173	181	188	195	203	210	217	225	
592	232	240	247	254	262	269	276	283	291	298	
593	305	313	320	327	335	342	349	357	364	371	
594	379	386	393	401	408	415	422	430	437	444	
595	452	459	466	474	481	488	495	503	510	517	
596	525	532	539	546	554	561	568	576	583	590	
597	597	605	612	619	627	634	641	648	656	663	
598	670	677	685	692	699	706	714	721	728	735	
599	743	750	757	764	772	779	786	793	801	808	
600	815	822	830	837	844	851	859	866	873	880	
N	**L 0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Prop. Pts.**

Appendix Table X

Five-Place Logarithms of Numbers

600–650

N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.
600	77 815	822	830	837	844	851	859	866	873	880	
601	887	895	902	909	916	924	931	938	945	952	
602	960	967	974	981	988	996	*003	*010	*017	*025	
603	78 032	039	046	053	061	068	075	082	089	097	
604	104	111	118	125	132	140	147	154	161	168	
605	176	183	190	197	204	211	219	226	233	240	
606	247	254	262	269	276	283	290	297	305	312	
607	319	326	333	340	347	355	362	369	376	383	8
608	390	398	405	412	419	426	433	440	447	455	1 0.8
609	462	469	476	483	490	497	504	512	519	526	2 1.6
610	533	540	547	554	561	569	576	583	590	597	3 2.4 4 3.2 5 4.0
611	604	611	618	625	633	640	647	654	661	668	6 4.8
612	675	682	689	696	704	711	718	725	732	739	7 5.6
613	746	753	760	767	774	781	789	796	803	810	8 6.4
614	817	824	831	838	845	852	859	866	873	880	9 7.2
615	888	895	902	909	916	923	930	937	944	951	
616	958	965	972	979	986	993	*000	*007	*014	*021	
617	79 029	036	043	050	057	064	071	078	085	092	
618	099	106	113	120	127	134	141	148	155	162	
619	169	176	183	190	197	204	211	218	225	232	
620	239	246	253	260	267	274	281	288	295	302	
621	309	316	323	330	337	344	351	358	365	372	
622	379	386	393	400	407	414	421	428	435	442	
623	449	456	463	470	477	484	491	498	505	511	7
624	518	525	532	539	546	553	560	567	574	581	1 0.7
625	588	595	602	609	616	623	630	637	644	650	2 1.4 3 2.1 4 2.8
626	657	664	671	678	685	692	699	706	713	720	5 3.5
627	727	734	741	748	754	761	768	775	782	789	6 4.2
628	796	803	810	817	824	831	837	844	851	858	7 4.9
629	865	872	879	886	893	900	906	913	920	927	8 5.6 9 6.3
630	934	941	948	955	962	969	975	982	989	996	
631	80 003	010	017	024	030	037	044	051	058	065	
632	072	079	085	092	099	106	113	120	127	134	
633	140	147	154	161	168	175	182	188	195	202	
634	209	216	223	229	236	243	250	257	264	271	
635	277	284	291	298	305	312	318	325	332	339	
636	346	353	359	366	373	380	387	393	400	407	
637	414	421	428	434	441	448	455	462	468	475	6
638	482	489	496	502	509	516	523	530	536	543	1 0.6
639	550	557	564	570	577	584	591	598	604	611	2 1.2
640	618	625	632	638	645	652	659	665	672	679	3 1.8 4 2.4 5 3.0
641	686	693	699	706	713	720	726	733	740	747	6 3.6
642	754	760	767	774	781	787	794	801	808	814	7 4.2
643	821	828	835	841	848	855	862	868	875	882	8 4.8
644	889	895	902	909	916	922	929	936	943	949	9 5.4
645	956	963	969	976	983	990	996	*003	*010	*017	
646	81 023	030	037	043	050	057	064	070	077	084	
647	090	097	104	111	117	124	131	137	144	151	
648	158	164	171	178	184	191	198	204	211	218	
649	224	231	238	245	251	258	265	271	278	285	
650	291	298	305	311	318	325	331	338	345	351	
N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.

Appendix Table X

Five-Place Logarithms of Numbers

650–700

N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.
650	81 291	298	305	311	318	325	331	338	345	351	
651	358	365	371	378	385	391	398	406	411	418	
652	425	431	438	445	451	458	465	471	478	485	
653	491	498	505	511	518	525	531	538	544	551	
654	558	564	571	578	584	591	598	604	611	617	
655	624	631	637	644	651	657	664	671	677	684	
656	690	697	704	710	717	723	730	737	743	750	
657	757	763	770	776	783	790	796	803	809	816	
658	823	829	836	842	849	856	862	869	875	882	
659	889	895	902	908	915	921	928	935	941	948	
660	954	961	968	974	981	987	994	*000	*007	*014	
661	82 020	027	033	040	046	053	060	066	073	079	7
662	086	092	099	105	112	119	125	132	138	145	1 0.7
663	151	158	164	171	178	184	191	197	204	210	2 1.4
664	217	223	230	236	243	249	256	263	269	276	3 2.1
665	282	289	295	302	308	315	321	328	334	341	4 2.8
666	347	354	360	367	373	380	387	393	400	406	5 3.5
667	413	419	426	432	439	445	452	458	465	471	6 4.2
668	478	484	491	497	504	510	517	523	530	536	7 4.9
669	543	549	556	562	569	575	582	588	595	601	8 5.6
											9 6.3
670	607	614	620	627	633	640	646	653	659	666	
671	672	679	685	692	698	705	711	718	724	730	
672	737	743	750	756	763	769	776	782	789	795	
673	802	808	814	821	827	834	840	847	853	860	
674	866	872	879	885	892	898	905	911	918	924	
675	930	937	943	950	956	963	969	975	982	988	
676	995	*001	*008	*014	*020	*027	*033	*040	*046	*052	
677	83 059	065	072	078	085	091	097	104	110	117	
678	123	129	136	142	149	155	161	168	174	181	
679	187	193	200	206	213	219	225	232	238	245	
680	251	257	264	270	276	283	289	296	302	308	
681	315	321	327	334	340	347	353	359	366	372	
682	378	385	391	398	404	410	417	423	429	436	6
683	442	448	455	461	467	474	480	487	493	499	1 0.6
684	506	512	518	525	531	537	544	550	556	563	2 1.2
685	569	575	582	588	594	601	607	613	620	626	3 1.8
686	632	639	645	651	658	664	670	677	683	689	4 2.4
687	696	702	708	715	721	727	734	740	746	753	5 3.0
688	759	765	771	778	784	790	797	803	809	816	6 3.6
689	822	828	835	841	847	853	860	866	872	879	7 4.2
											8 4.8
											9 5.4
690	885	891	897	904	910	916	923	929	935	942	
691	948	954	960	967	973	979	985	992	998	*004	
692	84 011	017	023	029	036	042	048	055	061	067	
693	073	080	086	092	098	105	111	117	123	130	
694	136	142	148	155	161	167	173	180	186	192	
695	198	205	211	217	223	230	236	242	248	255	
696	261	267	273	280	286	292	298	305	311	317	
697	323	330	336	342	348	354	361	367	373	379	
698	386	392	398	404	410	417	423	429	435	442	
699	448	454	460	466	473	479	485	491	497	504	
700	510	516	522	528	535	541	547	553	559	566	
N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.

Appendix Table X

Five-Place Logarithms of Numbers

700–750

N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.
700	84 310	516	522	528	535	541	547	553	559	566	
701	572	578	584	590	597	603	609	615	621	628	
702	634	640	646	652	658	665	671	677	683	689	
703	696	702	708	714	720	726	733	739	745	751	
704	757	763	770	776	782	788	794	800	807	813	
705	819	825	831	837	844	850	856	862	868	874	
706	880	887	893	899	905	911	917	924	930	936	7
707	942	948	954	960	967	973	979	985	991	997	
708	85 003	009	016	022	028	034	040	046	052	058	1 0.7
709	065	071	077	083	089	095	101	107	114	120	2 1.4
											3 2.1
710	126	132	138	144	150	156	163	169	175	181	4 2.8
711	187	193	199	205	211	217	224	230	236	242	5 3.5
712	248	254	260	266	272	278	285	291	297	303	6 4.2
713	309	315	321	327	333	339	345	352	358	364	7 4.9
714	370	376	382	388	394	400	406	412	418	425	8 5.6
											9 6.3
715	431	437	443	449	455	461	467	473	479	485	
716	491	497	503	509	516	522	528	534	540	546	
717	552	558	564	570	576	582	588	594	600	606	
718	612	618	625	631	637	643	649	655	661	667	
719	673	679	685	691	697	703	709	715	721	727	
720	733	739	745	751	757	763	769	775	781	788	
721	794	800	806	812	818	824	830	836	842	848	
722	854	860	866	872	878	884	890	896	902	908	6
723	914	920	926	932	938	944	950	956	962	968	1 0.6
724	974	980	986	992	998	*004	*010	*016	*022	*028	2 1.2
											3 1.8
725	86 034	040	046	052	058	064	070	076	082	088	4 2.4
726	094	100	106	112	118	124	130	136	141	147	5 3.0
727	153	159	165	171	177	183	189	195	201	207	6 3.6
728	213	219	225	231	237	243	249	255	261	267	7 4.2
729	273	279	285	291	297	303	308	314	320	326	8 4.8
											9 5.4
730	332	338	344	350	356	362	368	374	380	386	
731	392	398	404	410	415	421	427	433	439	445	
732	451	457	463	469	475	481	487	493	499	504	
733	510	516	522	528	534	540	546	552	558	564	
734	570	576	581	587	593	599	605	611	617	623	
735	629	635	641	646	652	658	664	670	676	682	
736	688	694	700	705	711	717	723	729	735	741	
737	747	753	759	764	770	776	782	788	794	800	5
738	806	812	817	823	829	835	841	847	853	859	1 0.5
739	864	870	876	882	888	894	900	906	911	917	2 1.0
											3 1.5
740	923	929	935	941	947	953	958	964	970	976	4 2.0
741	982	988	994	999	*005	*011	*017	*023	*029	*035	5 2.5
742	87 040	046	052	058	064	070	075	081	087	093	6 3.0
743	099	105	111	116	122	128	134	140	146	151	7 3.5
744	157	163	169	175	181	186	192	198	204	210	8 4.0
											9 4.5
745	216	221	227	233	239	245	251	256	262	268	
746	274	280	286	291	297	303	309	315	320	326	
747	332	338	344	349	355	361	367	373	379	384	
748	390	396	402	408	413	419	425	431	437	442	
749	448	454	460	466	471	477	483	489	495	500	
750	506	512	518	523	529	535	541	547	552	558	
N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.

Appendix Table X

Five-Place Logarithms of Numbers

750–800

N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.
750	87 506	512	518	523	529	535	541	547	552	558	
751	564	570	576	581	587	593	599	604	610	616	
752	622	628	633	639	645	651	656	662	668	674	
753	679	685	691	697	703	708	714	720	726	731	
754	737	743	749	754	760	766	772	777	783	789	
755	795	800	806	812	818	823	829	835	841	846	
756	852	858	864	869	875	881	887	892	898	904	
757	910	915	921	927	933	938	944	950	955	961	
758	967	973	978	984	990	996	*001	*007	*013	*018	
759	88 024	030	036	041	047	053	058	064	070	076	
760	081	087	093	098	104	110	116	121	127	133	
761	138	144	150	156	161	167	173	178	184	190	6
762	195	201	207	213	218	224	230	235	241	247	1 0.6
763	252	258	264	270	275	281	287	292	298	304	2 1.2
764	309	315	321	326	332	338	343	349	355	360	3 1.8
765	366	372	377	383	389	395	400	406	412	417	4 2.4
766	423	429	434	440	446	451	457	463	468	474	5 3.0
767	480	485	491	497	502	508	513	519	525	530	6 3.6
768	536	542	547	553	559	564	570	576	581	587	7 4.2
769	593	598	604	610	615	621	627	632	638	643	8 4.8
											9 5.4
770	649	655	660	666	672	677	683	689	694	700	
771	705	711	717	722	728	734	739	745	750	756	
772	762	767	773	779	784	790	795	801	807	812	
773	818	824	829	835	840	846	852	857	863	868	
774	874	880	885	891	897	902	908	913	919	925	
775	930	936	941	947	953	958	964	969	975	981	
776	986	992	997	*003	*009	*014	*020	*025	*031	*037	
777	89 042	048	053	059	064	070	076	081	087	092	
778	098	104	109	115	120	126	131	137	143	148	
779	154	159	165	170	176	182	187	193	198	204	
780	209	215	221	226	232	237	243	248	254	260	
781	265	271	276	282	287	293	298	304	310	315	
782	321	326	332	337	343	348	354	360	365	371	5
783	376	382	387	393	398	404	409	415	421	426	1 0.5
784	432	437	443	448	454	459	465	470	476	481	2 1.0
785	487	492	498	504	509	515	520	526	531	537	3 1.5
786	542	548	553	559	564	570	575	581	586	592	4 2.0
787	597	603	609	614	620	625	631	636	642	647	5 2.5
788	653	658	664	669	675	680	686	691	697	702	6 3.0
789	708	713	719	724	730	735	741	746	752	757	7 3.5
											8 4.0
											9 4.5
790	763	768	774	779	785	790	796	801	807	812	
791	818	823	829	834	840	845	851	856	862	867	
792	873	878	883	889	894	900	905	911	916	922	
793	927	933	938	944	949	955	960	966	971	977	
794	982	988	993	998	*004	*009	*015	*020	*026	*031	
795	90 037	042	048	053	059	064	069	075	080	086	
796	091	097	102	108	113	119	124	129	135	140	
797	146	151	157	162	168	173	179	184	189	195	
798	200	206	211	217	222	227	233	238	244	249	
799	255	260	266	271	276	282	287	293	298	304	
800	309	314	320	325	331	336	342	347	352	358	
N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.

Appendix Table X

Five-Place Logarithms of Numbers

800–850

N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.
800	90 309	314	320	325	331	336	342	347	352	358	
801	363	369	374	380	385	390	396	401	407	412	
802	417	423	428	434	439	445	450	455	461	466	
803	472	477	482	488	493	499	504	509	515	520	
804	526	531	536	542	547	553	558	563	569	574	
805	580	585	590	596	601	607	612	617	623	628	
806	634	639	644	650	655	660	666	671	677	682	
807	687	693	698	703	709	714	720	725	730	736	
808	741	747	752	757	763	768	773	779	784	789	
809	795	800	806	811	816	822	827	832	838	843	
810	849	854	859	865	870	875	881	886	891	897	
811	902	907	913	918	924	929	934	940	945	950	6
812	956	961	966	972	977	982	988	993	998	*004	
813	91 009	014	020	025	030	036	041	046	052	057	1 0.6
814	062	068	073	078	084	089	094	100	105	110	2 1.2
815	116	121	126	132	137	142	148	153	158	164	3 1.8
816	169	174	180	185	190	196	201	206	212	217	4 2.4
817	222	228	233	238	243	249	254	259	265	270	5 3.0
818	275	281	286	291	297	302	307	312	318	323	6 3.6
819	328	334	339	344	350	355	360	365	371	376	7 4.2
											8 4.8
											9 5.4
820	381	387	392	397	403	408	413	418	424	429	
821	434	440	445	450	455	461	466	471	477	482	
822	487	492	498	503	508	514	519	524	529	535	
823	540	545	551	556	561	566	572	577	582	587	
824	593	598	603	609	614	619	624	630	635	640	
825	645	651	656	661	666	672	677	682	687	693	
826	698	703	709	714	719	724	730	735	740	745	
827	751	756	761	766	772	777	782	787	793	798	
828	803	808	814	819	824	829	834	840	845	850	
829	855	861	866	871	876	882	887	892	897	903	
830	908	913	918	924	929	934	939	944	950	955	
831	960	965	971	976	981	986	991	997	*002	*007	
832	92 012	018	023	028	033	038	044	049	054	059	5
833	065	070	075	080	085	091	096	101	106	111	1 0.5
834	117	122	127	132	137	143	148	153	158	163	2 1.0
835	169	174	179	184	189	195	200	205	210	215	3 1.5
836	221	226	231	236	241	247	252	257	262	267	4 2.0
837	273	278	283	288	293	298	304	309	314	319	5 2.5
838	324	330	335	340	345	350	355	361	366	371	6 3.0
839	376	381	387	392	397	402	407	412	418	423	7 3.5
											8 4.0
											9 4.5
840	428	433	438	443	449	454	459	464	469	474	
841	480	485	490	495	500	505	511	516	521	526	
842	531	536	542	547	552	557	562	567	572	578	
843	583	588	593	598	603	609	614	619	624	629	
844	634	639	645	650	655	660	665	670	675	681	
845	686	691	696	701	706	711	716	722	727	732	
846	737	742	747	752	758	763	768	773	778	783	
847	788	793	799	804	809	814	819	824	829	834	
848	840	845	850	855	860	865	870	875	881	886	
849	891	896	901	906	911	916	921	927	932	937	
850	942	947	952	957	962	967	973	978	983	988	
N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.

Appendix Table X

Five-Place Logarithms of Numbers

850–900

N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.
850	92 942	947	952	957	962	967	973	978	983	988	
851	993	998	*003	*008	*013	*018	*024	*029	*034	*039	
852	93 044	049	054	059	064	069	075	080	085	090	
853	095	100	105	110	115	120	125	131	136	141	
854	146	151	156	161	166	171	176	181	186	192	
855	197	202	207	212	217	222	227	232	237	242	
856	247	252	258	263	268	273	278	283	288	293	
857	298	303	308	313	318	323	328	334	339	344	6
858	349	354	359	364	369	374	379	384	389	394	1 0.6
859	399	404	409	414	420	425	430	435	440	445	2 1.2
860	450	455	460	465	470	475	480	485	490	495	3 1.8 4 2.4 5 3.0
861	500	505	510	515	520	526	531	536	541	546	6 3.6
862	551	556	561	566	571	576	581	586	591	596	7 4.2
863	601	606	611	616	621	626	631	636	641	646	8 4.8
864	651	656	661	666	671	676	682	687	692	697	9 5.4
865	702	707	712	717	722	727	732	737	742	747	
866	752	757	762	767	772	777	782	787	792	797	
867	802	807	812	817	822	827	832	837	842	847	
868	852	857	862	867	872	877	882	887	892	897	
869	902	907	912	917	922	927	932	937	942	947	
870	952	957	962	967	972	977	982	987	992	997	
871	94 002	007	012	017	022	027	032	037	042	047	
872	052	057	062	067	072	077	082	086	091	096	5
873	101	106	111	116	121	126	131	136	141	146	1 0.5
874	151	156	161	166	171	176	181	186	191	196	2 1.0
875	201	206	211	216	221	226	231	236	240	245	3 1.5 4 2.0 5 2.5
876	250	255	260	265	270	275	280	285	290	295	6 3.0
877	300	305	310	315	320	325	330	335	340	345	7 3.5
878	349	354	359	364	369	374	379	384	389	394	8 4.0
879	399	404	409	414	419	424	429	433	438	443	9 4.5
880	448	453	458	463	468	473	478	483	488	493	
881	498	503	507	512	517	522	527	532	537	542	
882	547	552	557	562	567	571	576	581	586	591	
883	596	601	606	611	616	621	626	630	635	640	
884	645	650	655	660	665	670	675	680	685	689	
885	694	699	704	709	714	719	724	729	734	738	
886	743	748	753	758	763	768	773	778	783	787	
887	792	797	802	807	812	817	822	827	832	836	4
888	841	846	851	856	861	866	871	876	880	885	1 0.4
889	890	895	900	905	910	915	919	924	929	934	2 0.8
890	939	944	949	954	959	963	968	973	978	983	3 1.2 4 1.6 5 2.0
891	988	993	998	*002	*007	*012	*017	*022	*027	*032	6 2.4
892	95 036	041	046	051	056	061	066	071	075	080	7 2.8
893	085	090	095	100	105	109	114	119	124	129	8 3.2
894	134	139	143	148	153	158	163	168	173	177	9 3.6
895	182	187	192	197	202	207	211	216	221	226	
896	231	236	240	245	250	255	260	265	270	274	
897	279	284	289	294	299	303	308	313	318	323	
898	328	332	337	342	347	352	357	361	366	371	
899	376	381	386	390	395	400	405	410	415	419	
900	424	429	434	439	444	448	453	458	463	468	
N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.

Appendix Table X

Five-Place Logarithms of Numbers

900–950

N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.
900	95 424	429	434	439	444	448	453	458	463	468	
901	472	477	482	487	492	497	501	506	511	516	
902	521	525	530	535	540	545	550	554	559	564	
903	569	574	578	583	588	593	598	602	607	612	
904	617	622	626	631	636	641	646	650	655	660	
905	665	670	674	679	684	689	694	698	703	708	
906	713	718	722	727	732	737	742	746	751	756	
907	761	766	770	775	780	785	789	794	799	804	
908	809	813	818	823	828	832	837	842	847	852	
909	856	861	866	871	875	880	885	890	895	899	
910	904	909	914	018	923	928	933	938	942	947	
911	952	957	961	966	971	976	980	985	990	995	5
912	999	*004	*009	*014	*019	*023	*028	*033	*038	*042	1 0.5
913	96 047	052	057	061	066	071	076	080	085	090	2 1.0
914	095	099	104	109	114	118	123	128	133	137	3 1.5
915	142	147	152	156	161	166	171	175	180	185	4 2.0
916	190	194	199	204	209	213	218	223	227	232	5 2.5
917	237	242	246	251	256	261	265	270	275	280	6 3.0
918	284	289	294	298	303	308	313	317	322	327	7 3.5
919	332	336	341	346	350	355	360	365	369	374	8 4.0
											9 4.5
920	379	384	388	393	398	402	407	412	417	421	
921	426	431	435	440	445	450	454	459	464	468	
922	473	478	483	487	492	497	501	506	511	515	
923	520	525	530	534	539	544	548	553	558	562	
924	567	572	577	581	586	591	595	600	605	609	
925	614	619	624	628	633	638	642	647	652	656	
926	661	666	670	675	680	685	689	694	699	703	
927	708	713	717	722	727	731	736	741	745	750	
928	755	759	764	769	774	778	783	788	792	797	
929	802	806	811	816	820	825	830	834	839	844	
930	848	853	858	862	867	872	876	881	886	890	
931	895	900	904	909	914	918	923	928	932	937	
932	942	946	951	956	960	965	970	974	979	984	4
933	988	993	997	*002	*007	*011	*016	*021	*025	*030	1 0.4
934	97 035	039	044	049	053	058	063	067	072	077	2 0.8
935	081	086	090	095	100	104	109	114	118	123	3 1.2
936	128	132	137	142	146	151	155	160	165	169	4 1.6
937	174	179	183	188	192	197	202	206	211	216	5 2.0
938	220	225	230	234	239	243	248	253	257	262	6 2.4
939	267	271	276	280	285	290	294	299	304	308	7 2.8
											8 3.2
											9 3.6
940	313	317	322	327	331	336	340	345	350	354	
941	359	364	368	373	377	382	387	391	396	400	
942	405	410	414	419	424	428	433	437	442	447	
943	451	456	460	465	470	474	479	483	488	493	
944	497	502	506	511	516	520	525	529	534	539	
945	543	548	552	557	562	566	571	575	580	585	
946	589	594	598	603	607	612	617	621	626	630	
947	635	640	644	649	653	658	663	667	672	676	
948	681	685	690	695	699	704	708	713	717	722	
949	727	731	736	740	745	749	754	759	763	768	
950	772	777	782	786	791	795	800	804	809	813	
N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.

APPENDIX TABLE X

Five-Place Logarithms of Numbers

950–1000

N	L 0	1	2	3	4	5	6	7	8	9	Prop. Pts.
950	97 772	777	782	786	791	795	800	804	809	813	
951	818	823	827	832	836	841	845	850	855	859	
952	864	868	873	877	882	886	891	896	900	905	
953	909	914	918	923	928	932	937	941	946	950	
954	955	959	964	968	973	978	982	987	991	996	
955	98 000	005	009	014	019	023	028	032	037	041	
956	046	050	055	059	064	068	073	078	082	087	
957	091	096	100	105	109	114	118	123	127	132	
958	137	141	146	150	155	159	164	168	173	177	
959	182	186	191	195	200	204	209	214	218	223	
960	227	232	236	241	245	250	254	259	263	268	
961	272	277	281	286	290	295	299	304	308	313	5
962	318	322	327	331	336	340	345	349	354	358	
963	363	367	372	376	381	385	390	394	399	403	1 0.5
964	408	412	417	421	426	430	435	439	444	448	2 1.0
965	453	457	462	466	471	475	480	484	489	493	3 1.5
966	498	502	507	511	516	520	525	529	534	538	4 2.0
967	543	547	552	556	561	565	570	574	579	583	5 2.5
968	588	592	597	601	605	610	614	619	623	628	6 3.0
969	632	637	641	646	650	655	659	664	668	673	7 3.5
											8 4.0
											9 4.5
970	677	682	686	691	695	700	704	709	713	717	
971	722	726	731	735	740	744	749	753	758	762	
972	767	771	776	780	784	789	793	798	802	807	
973	811	816	820	825	829	834	838	843	847	851	
974	856	860	865	869	874	878	883	887	892	896	
975	900	905	909	914	918	923	927	932	936	941	
976	945	949	954	958	963	967	972	976	981	985	
977	989	994	998	*003	*007	*012	*016	*021	*025	*029	
978	99 034	038	043	047	052	056	061	065	069	074	
979	078	083	087	092	096	100	105	109	114	118	
980	123	127	131	136	140	145	149	154	158	162	
981	167	171	176	180	185	189	193	198	202	207	
982	211	216	220	224	229	233	238	242	247	251	4
983	255	260	264	269	273	277	282	286	291	295	
984	300	304	308	313	317	322	326	330	335	339	1 0.4
985	344	348	352	357	361	366	370	374	379	383	2 0.8
986	388	392	396	401	405	410	414	419	423	427	3 1.2
987	432	436	441	445	449	454	458	463	467	471	4 1.6
988	476	480	484	489	493	498	502	506	511	515	5 2.0
989	520	524	528	533	537	542	546	550	555	559	6 2.4
											7 2.8
											8 3.2
											9 3.6
990	564	568	572	577	581	585	590	594	599	603	
991	607	612	616	621	625	629	634	638	642	647	
992	651	656	660	664	669	673	677	682	686	691	
993	695	699	704	708	712	717	721	726	730	734	
994	739	743	747	752	756	760	765	769	774	778	
995	782	787	791	795	800	804	808	813	817	822	
996	826	830	835	839	843	848	852	856	861	865	
997	870	874	878	883	887	891	896	900	904	909	
998	913	917	922	926	930	935	939	944	948	952	
999	957	961	965	970	974	978	983	987	991	996	
1000	00 000	004	009	013	017	022	026	030	035	039	
N	**L 0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Prop. Pts.**

LIST OF REFERENCES

American Recommended Practice, Engineering and Scientific Charts for Lantern Slides. New York, American Society of Mechanical Engineers.

ANDERSON, OSKAR N., "Statistical Method." *Encyclopaedia of the Social Sciences*, Vol. 14.

ANGELL, JAMES W., *The Behavior of Money.* New York, McGraw Hill, 1936.

ARKIN H. and COLTON, R. R., *Graphs: How to Make and Use Them.* New York, Harper, 1936.

BARLOW, *Tables of Squares, Cubes, Square Roots, Cube Roots and Reciprocals.* New York, Spar and Chamberlain.

BEAN, L. H., *Application of a Simplified Method of Graphic Curvilinear Correlation.* Bureau of Agricultural Economics, U. S. Department of Agriculture, 1929. "A Simplified Method of Graphic Curvilinear Correlation." *Journal of the American Statistical Association*, Dec., 1929.

BECKETT, S. H. and ROBERTSON, R. D., *The Economical Irrigation of Alfalfa in the Sacramento Valley.* Bulletin 280, Agricultural Experiment Station, University of California, May, 1917.

BENEY, M. ADA, *Cost of Living in the United States, 1914–1936.* New York, National Industrial Conference Board, 1936.

BLISS, CHARLES A., *Production in Depression and Recovery.* Bulletin 58, National Bureau of Economic Research, Nov., 1935.

BODDINGTON, A. L., *Statistics and Their Application to Commerce.* London, Sir Isaac Pitman and Sons, 1934.

BOWLEY, A. L., *Elements of Statistics.* London, P. S. King and Son, 1937. "On the Precision Attained in Sampling." *Bulletin*, International Statistical Institute, 1926. "The Application of Sampling to Economic and Sociological Problems." *Journal of the American Statistical Association*, Sept., 1936.

BRINTON, W. C., *Graphic Methods of Presenting Facts.* New York, The Engineering Magazine Co., 1914.

BROAD, C. D., "On the Relation between Induction and Probability." *Mind*, N. S. Vol. 27, 1918, and Vol. 29, 1920.

BRUNT, DAVID, *The Combination of Observations.* Cambridge University Press, 1917.

BURGESS, ROBERT W., *Introduction to the Mathematics of Statistics.* Boston, Houghton Mifflin, 1927.

BURNS, ARTHUR F., *Production Trends in the United States since 1870.* New York. National Bureau of Economic Research, 1928. "The Measurement of the Physical Volume of Production." *Quarterly Journal of Economics,* Feb., 1930.

CAMP, B. H., *The Mathematical Part of Elementary Statistics.* New York, D. C. Heath and Co., 1934.

CARVER, H. C., "Frequency Curves." In *Handbook of Mathematical Statistics,* Rietz, H. L. ed. Boston, Houghton Mifflin, 1924.

CHADDOCK, R. E., *Principles and Methods of Statistics.* Boston, Houghton Mifflin, 1925.

CHARLIER, C. V. L., *Vorlesungen Über Die Grundzüge Der Mathematischen Statistik.* Lund, Verlag Scientia, 1920.

CLARK, WALLACE, *The Gantt Chart.* New York, Ronald Press, 1922.

Code of Preferred Practice for Graphic Presentation: Time Series Charts. New York, American Society of Mechanical Engineers, 1936.

COHEN, MORRIS R., "The Statistical View of Nature." *Journal of the American Statistical Association,* June, 1936.

CROXTON, F. E. and COWDEN, D. J., *Practical Business Statistics.* New York, Prentice Hall, 1934.

CRUM, W. L. and PATTON, A. C., *An Introduction to the Methods of Economic Statistics.* New York, A. W. Shaw Co., 1925.

CUTTS, JESSE M. and DENNIS, SAMUEL J., "Revised Method of Calculation of the B. L. S. Wholesale Price Index." *Journal of the American Statistical Association,* Dec., 1937.

DAVENPORT, DONALD H. and SCOTT, FRANCES V., *An Index to Business Indices.* Chicago, Business Publications, Inc., 1937.

DAVENPORT, E., "Comparative Agriculture." In *Bailey's Cyclopedia of American Agriculture.*

DAVIES, GEORGE R. and CROWDER, W. F., *Methods of Statistical Analysis in the Social Sciences.* New York, Wiley, 1933.

DAVIES, GEORGE R. and YODER, DALE, *Business Statistics.* New York, Wiley, 1937.

DAVIS, HAROLD T. and NELSON, W. F. C., *Elements of Statistics.* Bloomington, Ind., The Principia Press, Inc., 1935.

DAY, E. E., "An Index of the Physical Volume of Production." *Review of Economic Statistics,* Sept., 1920, Jan., 1921. *Statistical Analysis.* New York, Macmillan, 1925.

DAY, E. E. and THOMAS, W., *The Growth of Manufactures, 1899–1923.* Census Monographs, VIII, 1928.

DEMING, W. EDWARDS and BIRGE, RAYMOND T., *On the Statistical Theory of Errors.* Reprint from *Reviews of Modern Physics,* Vol. 6, July, 1934. Washington, Graduate School, U. S. Department of Agriculture.

Editorial, "On the Probable Errors of Frequency Constants." *Biometrika,* Vol. 2.

ELDERTON, W. P., *Frequency Curves and Correlation.* London, Layton, 1906.

EZEKIEL, MORDECAI, "A Method of Handling Curvilinear Correlation for Any Number of Variables." *Journal of the American Statistical Association,* Vol. 19, N. S. No. 148, 1924. "Correlation." *Encyclopaedia of the Social Sciences,* Vol. 4. *Methods of Correlation Analysis.* New York, Wiley, 1930.

FALKNER, HELEN D., "The Measurement of Seasonal Variation." *Journal of the American Statistical Association,* June, 1924.

FERGER, WIRTH F., "Distinctive Concepts of Price and Purchasing-Power Index Numbers." *Journal of the American Statistical Association,* Vol. 31, No. 194, June, 1936.

FISHER, ARNE, *An Elementary Treatise on Frequency Curves.* New York, Macmillan, 1922. *The Mathematical Theory of Probabilities.* New York, Macmillan, 1922.

FISHER, IRVING, *The Making of Index Numbers.* Boston, Houghton Mifflin, 1922.

FISHER, R. A., *Statistical Methods for Research Workers.* Edinburgh, Oliver and Boyd, 6th ed., 1936. *The Design of Experiments.* Edinburgh, Oliver and Boyd, 1935. "The Mathematical Distributions Used in the Common Tests of Significance." *Econometrica,* Vol. 3, No. 4.

FLORENCE, P. S., *The Statistical Method in Economics and Political Science.* New York, Harcourt Brace and Co., 1929.

FLUX, A. W., "The Measurement of Price Changes." *Journal of the Royal Statistical Society,* March, 1921.

FORSYTH, C. H., *An Introduction to the Mathematical Analysis of Statistics.* New York, Wiley, 1924.

FRAZIER, EDWARD K., "Earnings and Hours in Blast Furnaces, Bessemer Converters, Open-Hearth Furnaces and Electric Furnaces, 1933 and 1935." *Monthly Labor Review,* April, 1936.

FRICKEY, EDWIN, "The Problem of Secular Trend." *Review of Economic Statistics,* Oct., 1934.

FRIEDMAN, MILTON, "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance." *Journal of the American Statistical Association*, Vol. 32, Dec., 1937.

FRY, THORNTON C., *Probability and Its Engineering Uses.* New York, Van Nostrand, 1928.

GALTON, FRANCIS, "Correlations and Their Measurement." *Proceedings of the Royal Society*, Vol. 45, 1888.

GLOVER, JAMES W., *Tables of Applied Mathematics.* Ann Arbor, Michigan, George Wahr, 1923.

GRIFFIN, F. L., *Introduction to Mathematical Analysis.* Boston, Houghton Mifflin, 1922.

HAAS, G. C., *Sale Prices as a Basis for Farm Land Appraisal.* Technical Bulletin No. 9, University of Minnesota Agricultural Experiment Station.

HABERLER, G., *Der Sinn der Indexzahlen.* Tübingen, J. C. B. Mohr, 1927.

HALL, LINCOLN W., "Seasonal Variation as a Relative of Secular Trend." *Journal of the American Statistical Association*, June, 1924.

HART, HORNELL, "The Reliability of a Percentage." *Journal of the American Statistical Association*, Vol. 21, Mar., 1926.

HASKELL, A. C., *How to Make and Use Graphic Charts.* New York, Codex Book Co., 1922.

HOTELLING, HAROLD, "The Generalization of Student's Ratio." *Annals of Mathematical Statistics*, Vol. 2, 1931.

HOTELLING, HAROLD and PABST, MARGARET, "Rank Correlation and Tests of Significance Involving No Assumption of Normality." *Annals of Mathematical Statistics*, Vol. 7, 1936.

HUNTINGTON, E. V., *Curve Fitting by the Method of Least Squares and the Method of Moments.* In *Handbook of Mathematical Statistics*, Rietz, H. L. ed. Boston, Houghton Mifflin, 1924.

JEROME, HARRY, *Statistical Method.* New York, Harper, 1924.

JONES, D. C., *A First Course in Statistics.* London, Bell, 1921.

KARSTEN, KARL G., *Charts and Graphs.* New York, Prentice Hall, 1923.

KELLEY, TRUMAN L., "Partial and Multiple Correlation." In *Handbook of Mathematical Statistics*, Rietz, H. L. ed. Boston, Houghton Mifflin, 1924. *Statistical Method.* New York, Macmillan, 1923.

KEYNES, J. M., *A Treatise on Money.* New York, Macmillan, 1930. *A Treatise on Probability.* New York, Macmillan, 1921.

KILLOUGH, HUGH B., *What Makes the Price of Oats?* Bulletin No. 1351, U. S. Department of Agriculture.

KING, W. I., *Elements of Statistical Method.* New York, Macmillan, 1912. *Index Numbers Elucidated.* New York, Longmans, Green & Co., 1930.

KNIBBS, SIR GEORGE, "The Nature of an Unequivocal Price-Index and Quantity-Index." *Journal of the American Statistical Association,* March, June, 1924.

KURTZ, EDWIN, "Replacement Insurance." *Administration,* July, 1921.

KUZNETS, SIMON, "On Moving Correlation of Time Sequences." *Journal of the American Statistical Association,* June, 1928. "Time Series." *Encyclopaedia of the Social Sciences,* Vol. 14.

LEONG, Y. S., "Indexes of the Physical Volume of Production of Producers' Goods, Consumers' Goods, Durable Goods and Transient Goods." *Journal of the American Statistical Association,* June, 1935.

LIPKA, JOSEPH, *Graphical and Mechanical Computation.* New York, Wiley, 1918.

LOVITT, W. V. and HOLTZCLAW, H. F., *Statistics.* New York, Prentice Hall, 1927.

MACAULAY, FREDERICK R., *The Smoothing of Time Series.* New York, National Bureau of Economic Research, 1931.

MALENBAUM, WILFRED and BLACK, JOHN D., "The Use of the Short-Cut Graphic Method of Multiple Correlation." *Quarterly Journal of Economics,* Nov., 1937.

MERRIMAN, MANSFIELD, *The Method of Least Squares.* New York, Wiley, 1897.

MILLS, F. C., "An Hypothesis Concerning the Duration of Business Cycles." *Journal of the American Statistical Association,* Dec., 1926. *Economic Tendencies in the United States.* New York, National Bureau of Economic Research, 1932. "On Measurement in Economics." In *The Trend of Economics,* Tugwell, R. G. ed. New York, Knopf, 1924. *Prices in Recession and Recovery.* New York, National Bureau of Economic Research, 1936. *The Behavior of Prices.* New York, National Bureau of Economic Research, 1927.

MILLS, F. C. and DAVENPORT, DONALD H., *Manual of Problems and Tables in Statistics.* New York, Holt, 1925.

MINER, J. R., *Tables of* $\sqrt{1-r^2}$ *and* $1-r^2$ *for use in Partial Correlation and Trigonometry.* Baltimore, Johns Hopkins Press, 1922.

MISES, R. VON, "Probability." *Encyclopaedia of the Social Sciences,* Vol. 12.

MITCHELL, W. C., *Business Cycles, The Problem and Its Setting.* New York, National Bureau of Economic Research, 1927. *The Making and Using of Index Numbers.* Part I. Bulletin 284, U. S. Bureau of Labor Statistics, Oct., 1921.

MITCHELL, W. C. and BURNS, ARTHUR F., *Production during the American Business Cycle of 1927–1933.* Bulletin 61, National Bureau of Economic Research, Nov., 1936.

MOORE, H. L., *Economic Cycles: Their Law and Cause.* New York, Macmillan, 1914. "Elasticity of Demand and Flexibility of Prices." *Journal of the American Statistical Association,* March, 1922. "Empirical Laws of Demand and Supply and the Flexibility of Prices." *Political Science Quarterly,* Dec., 1919. *Forecasting the Yield and the Price of Cotton.* New York, Macmillan, 1917. *Generating Economic Cycles.* New York, Macmillan, 1923.

MUDGETT, BRUCE D., *Statistical Tables and Graphs.* Boston, Houghton Mifflin, 1930.

NAGEL, ERNEST, "The Meaning of Probability." *Journal of the American Statistical Association,* March, 1936.

National Bureau of Economic Research. *Income in the United States,* Mitchell, W. C. ed. New York, Harcourt Brace and Co., 1921.

NEYMAN, J. (With the editorial assistance of W. E. Deming), *Lectures and Conferences on Mathematical Statistics.* Washington, Graduate School, U. S. Department of Agriculture, 1937.

PEAKE, E. G., *An Academic Study of Some Money Market and Other Statistics.* London, P. S. King, 1923.

PEARL, RAYMOND, *Introduction to Medical Biometry and Statistics.* Philadelphia, Saunders, 1923.

PEARSON, F. S., "Sampling Problems in Industry." *Journal of the Royal Statistical Society. Supplement,* Vol. 1, No. 2, 1934.

PEARSON, KARL, *Mathematical Contributions to the Theory of Evolution: On the General Theory of Skew Correlation and Non-Linear Regression.* Draper's Company Research Memoirs. Cambridge University Press, 1905. "Notes on the History of

Correlation." *Biometrika*, Vol. 13, 1920. "On a Correction Needful in the Case of the Correlation Ratio." *Biometrika*, Vol. 8, 1911. "On the Correction Necessary for the Correlation Ratio." *Biometrika*, Vol. 14, 1923. "On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be Reasonably Supposed to have Arisen from Random Sampling." *Philosophical Magazine*, 5th Series, Vol. 50, 1900. "Regression, Heredity and Panmixia. *Philosophical Transactions, Royal Society, Series A*, Vol. 187, 1896. *Tables for Statisticians and Biometricians.* London, Biometric Laboratory, University College, 2nd ed., 1924. "The Fundamental Problem of Practical Statistics." *Biometrika*, Vol. 13. *The Grammar of Science.* London, Black, 1911.

PERRY, F. G. and SILVERMAN, A. G., "A New Index of the Physical Volume of Canadian Business." *Journal of the American Statistical Association*, June, 1929.

PERSONS, WARREN M., "Correlation of Time Series." *Journal of the American Statistical Association*, June, 1923. "Fisher's Formula for Index Numbers." *Review of Economic Statistics*, Prel. Vol. III. *Forecasting Business Cycles.* New York, Wiley, 1931. "Indices of Business Conditions." *Review of Economic Statistics.* Prel. Vol. I, 1919. "The Variate Difference Correlation Method and Curve Fitting." *Quarterly Publications of the American Statistical Association*, June, 1917.

RHODES, E. C., *Elementary Statistical Methods.* London, George Routledge and Sons, 1933.

RICHARDSON, C. H., *An Introduction to Statistical Analysis.* New York, Harcourt Brace and Co., 1934.

RIDER, P. R., "A Survey of the Theory of Small Samples." *Annals of Mathematics*, Vol. 31, 1930.

RIEGEL, ROBERT, *Elements of Business Statistics.* New York, Appleton, 1927.

RIETZ, H. L. ed., *A Handbook of Mathematical Statistics.* Boston, Houghton Mifflin, 1924. "Random Sampling." In *A Handbook of Mathematical Statistics*, Rietz, H. L. ed. Boston, Houghton Mifflin, 1924.

RIGGLEMAN, JOHN R., *Graphic Methods for Presenting Business Statistics.* New York, McGraw Hill, 1936.

RIGGLEMAN, JOHN R. and FRISBEE, IRA N., *Business Statistics.* New York, McGraw Hill, 1932.

ROOS, CHARLES F., "The Correlation and Analysis of Time Series." *Econometrica*, Vol. 4, No. 4.

RUNNING, T. R., *Empirical Formulas*. New York, Wiley, 1917.

SARLE, CHARLES F., *Reliability and Adequacy of Farm Price Data*. Bulletin 1480, U. S. Department of Agriculture.

SASULY, MAX, *Trend Analysis of Statistics: Theory and Technique*. Washington, Brookings Institute, 1934.

SCHULTZ, HENRY, *Statistical Laws of Demand and Supply with Special Application to Sugar*. Chicago University Press, 1928. "The Standard Error of a Forecast from a Curve." *Journal of the American Statistical Association*, Vol. 25, No. 170.

SCHULTZ, T. W. and SNEDECOR, G. W., "Analysis of Variance as an Effective Method of Handling the Time Element in Certain Economic Statistics." *Journal of the American Statistical Association*, March, 1933.

SCHULTZE, ARTHUR, *Graphic Algebra*. New York, Macmillan, 1918.

SECRIST, HORACE, *Introduction to Statistical Methods*. New York, Macmillan, 1917.

SHEPPARD, W. F., "On the Calculation of the Most Probable Values of Frequency Constants for Data Arranged According to Equi-Distant Divisions of a Scale." *Proceedings of the London Mathematical Society*, Vol. 29, 1898. "The Calculation of the Moments of a Frequency Distribution." *Biometrika*, Vol. 5.

SHEWART, W. A., *Economic Control of Quality of Manufactured Product*. New York, Van Nostrand, 1931.

SMITH, B. B., "Combining the Advantages of First-difference and Deviation-from-Trend Methods of Correlating Time Series." *Journal of the American Statistical Association*, Vol. 21, 1926.

SMITH, J. G., *Elementary Statistics*. New York, Holt, 1934.

SNEDECOR, G. W., *Calculation and Interpretation of Analysis of Variance and Covariance*. Ames, Iowa, Collegiate Press, 1934. *Statistical Methods Applied to Experiments in Agriculture and Biology*. Ames, Iowa, Collegiate Press, 1937.

SNOW, E. C., "Trade Forecasting and Prices," *Journal of the Royal Statistical Society*, May, 1923.

SNYDER, CARL, *Business Cycles and Business Measurements*. New York, Macmillan, 1927.

SPURR, WILLIAM A., "A Graphic Method of Measuring Seasonal Variation," *Journal of the American Statistical Association*, June, 1937.

STAEHLE, H., "A Development of the Economic Theory of Price Index Numbers." *Review of Economic Studies*, June, 1935.

International Comparisons of Food Costs. Appendix I. In *Studies and Reports of the International Labor Office,* Series N, No. 20.

STAMP, J. C., "The Effect of Trade Fluctuations upon Profits." *Journal of the Royal Statistical Society,* July, 1918.

STEINMETZ, C. P., *Engineering Mathematics.* New York, McGraw Hill, 1917.

STEWART, ETHELBERT, "Labor Efficiency and Productiveness in Sawmills." *Monthly Labor Review,* Jan., 1923.

"Student," "The Probable Error of the Mean." *Biometrika,* Vol. 6, 1908.

STURGES, H. A., "The Choice of a Class Interval." *Journal of the American Statistical Association,* March, 1926.

SUTCLIFFE, W. G., *Elementary Statistical Methods.* New York, McGraw Hill, 1925.

THOMPSON, F. L., *Agricultural Prices.* New York, McGraw Hill, 1936.

TIPPETT, L. H. C., *The Methods of Statistics.* London, Williams and Norgate, 1937.

TOLLEY, H. R. and EZEKIEL, MORDECAI, "A Method of Handling Multiple Correlation Problems." *Journal of the American Statistical Association,* Dec., 1923.

WALKER, HELEN M., *Studies in the History of Statistical Method.* Baltimore, Williams and Wilkins, 1929.

WALSH, C. M., *The Measurement of General Exchange Value.* New York, Macmillan, 1901. *The Problem of Estimation.* London, P. S. King, 1921.

WARREN, G. F. and PEARSON, F. A., *Interrelationships of Supply and Price.* Ithaca, New York, Cornell University Agricultural Experiment Station, 1927.

WAUGH, A. E., *Elements of Statistical Method.* New York, McGraw Hill, 1938.

WEINTRAUB, DAVID, "Unemployment and Increasing Productivity." In *Technological Trends and National Policy,* Washington, National Resources Committee, June, 1937. (75th Congress, 1st Session, House Document No. 360.)

WELD, L. D., *Theory of Errors and Least Squares.* New York, Macmillan, 1916.

WHIPPLE, G. C., *Vital Statistics.* New York, Wiley, 1919.

WHITTAKER, E. T. and ROBINSON, G., *The Calculus of Observations.* London, Blackie and Son, 1924.

WILKS, S. S., "Test Criteria for Statistical Hypotheses Involving Several Variables." *Journal of the American Statistical Association*, Vol. 30, 1935. *The Theory of Statistical Inference, 1936–1937*. Planographed by Edwards Brothers, Ann Arbor, Michigan, 1937.

WILSON, E. B., "The Statistical Significance of Experimental Data." *Science*, Vol. 58, 1923.

WORKING, HOLBROOK, *Factors Determining the Price of Potatoes in St. Paul and Minneapolis*. Technical Bulletin 10, University of Minnesota Agricultural Experiment Station.

WORKING, HOLBROOK and HOTELLING, HAROLD, "The Application of the Theory of Error to the Interpretation of Trends." *Proceedings of the American Statistical Association*, March, 1929.

WRIGHT, T. W. and HAYFORD, J. F., *Adjustment of Observations*. New York, Van Nostrand, 1906.

YULE, G. U., "On the Time Correlation Problem, with Especial Reference to the Variate Difference Correlation Method." *Journal of the Royal Statistical Society*, July, 1921. "Table of the Values of P for Divergence from Independence in the Fourfold Table." *Journal of the Royal Statistical Society*, Vol. 85, Jan., 1922. "Why do we Sometimes get Nonsense Correlations between Time Series? A Study in Sampling and the Nature of Time Series." *Journal of the Royal Statistical Society*, Vol. 89, 1926.

YULE, G. U. and KENDALL, M. G., *An Introduction to the Theory of Statistics*. London, Charles Griffin and Co., 1937.

ZIZEK, FRANZ, *Statistical Averages*. New York. Holt. 1913.

INDEX

Abscissa, 9, 64
Accuracy of estimation, 332
Actuarial science, 272, 671
Aggregate, 307; use in index numbers, 165, 181; weighted, 193, 316
Alfalfa yield, correlation with irrigation, 404 ff.
American index numbers of wholesale prices, Bradstreet's index, 166, 182, 209, 211; Harvard index, 210; U. S. Bureau of Labor Statistics, 168, 172, 176, 193, 216 ff.
American Telephone and Telegraph Co., index of industrial activity, 312, 390, 393; study of frequency of telephone use, 440
Amplitude of cycles, 238, 285
Analysis of variance, *see* Variance analysis
Anderson, Oskar, 454, 487
Angell, James W., 270
Anti-logarithm, 24
Arbitrary origin, 351; in computing the mean, 106
Areas of the normal curve, 436 ff., 699
Arithmetic mean, 102; computation of, 103 ff.; weighted, 104; characteristics of, 126, 134; as most probable value, 332; moments about, 442; of the binomial distribution, 433, 660; significance of the difference between means, 481 ff.; significance of, in small samples, 605 ff.; use in correlation analysis, 537; standard error of, 464 ff., 664
Arithmetic series, 16, 28, 275
Array, 52, 82; in computing the correlation coefficient, 345
Artillery observations, 90
Astronomical observations, 89
Asymmetry, *see* Skewness
Average, 86 ff., 99, 101 ff.; relations between the several averages, 133; moving average, 234 ff.; of ratios to trend, 287
Average of relative prices, 183 ff., 196 ff.
Average relationship between variables, 328

Bank clearings, as index of business conditions, 242
Bar diagram, *see* Column diagram
Barlow's tables, 133
Base period, 316
Bean, Louis H., 564
Beckett, S. H., 405
Beta coefficients, 561
Bias, of index numbers, 191, 195; of the correlation index, 412; in sampling, 461, 599, 611
Binomial distribution, 429 ff., derivation of the mean and standard deviation of, 660
Birge, Raymond T., 469, 629
Black, John D., 564
Blakeman, John, 478
Bowley, A. L., 159, 201; representative sampling, 461; standard error, 472
Bradstreet's index, 166, 182, 209; use in deflation, 383
Burns, Arthur F., 242, 308
Business, 1; classes of activity, 1; quantitative character of its problems, 3, 5
Business cycle, 293 ff.; as indicated by moving averages, 242; as measured by production changes, 305; prewar relation to stock-price cycles, 390 ff.; post-war relation, 396; duration of, 467, 481; effect on price, 475; *see also* Cyclical variation

Census of manufactures, 309, 317, 371
Center of gravity, 102
Central tendency, 97, 99; measures of, 101 ff.
Certainty, in probability theory, 426
Chaddock, R. E., 84

Chain index, 216
Chance, law of, *see* Normal law of error and Probability
Characteristic, logarithmic, 24
Charlier check, 149
Charts, construction of, 32 ff.; for comparison of frequencies, 41 ff.; representation of component parts, 42 ff.; cumulative, 44 ff.; Gannt progress chart, 46; *see also* Graphic presentation
Chi-square, 618; distribution of, 618, 626; in testing goodness of fit, 626 ff.; in testing homogeneity, 633 ff.; in testing independence of principles of classification, 630 ff.; table of values, 625, 703
Classification of quantitative material, *see* Organization of data
Classification, principles of, 53 ff.; testing significance of, 494 ff.; testing independence of, 630 ff., 681 ff.
Classified data, *see* Grouping of data
Class interval, 53, 57 ff., 104, 347, 357; in locating the mode, 117; in computing the standard deviation, 149
Coefficient of correlation, *see* Correlation coefficient
Coefficient of multiple correlation, *see* Correlation, multiple
Coefficient of regression, *see* Regression
Coefficient of variation, 156
Coin tossing, 91
Column diagram, 41, 64 ff., 66, 73, 91
Commodities, included in price-change study, 209
Compound interest, law of, 30 ff., 40; curve of, 267
Concurrence of cycles, 390 ff.
Constants, 12, 244 ff.
Controls, in sampling procedure, 462
Coördinate geometry, 8 ff.
Correction, of index numbers, 311; of the correlation index, 412; of the standard error, 542
Correction factor, in computing the correlation coefficient, 339; in computing the mean, 106, 351, 392; *see also* Bias
Correlation, coefficient of, 334 ff., 520; calculation of, 337 ff., 353, 364, 648; product-moment method, 349 ff.; construction of table, 340 ff.; summary of correlation procedure, 366 ff.; limitations of, 370; relation to correlation ratio, 422; tests for the significance of, 502 ff., 611; significance of difference of coefficients, 616; standard error of, 474; derived from small samples, 610; weighted average of, 617, 618; table of significant values, 612, 701; table of relations to the *z* function, 702
Correlation, index of, 408 ff., 520; formula for, 408, 412, 647; significance of, 409; computation of, 410, 412; standard error, 477
Correlation, linear, 325 ff.; lines of regression, 359 ff.; distortion in non-normal distributions, 372 ff.; of grouped data, 340 ff.; in the measurement of time sequence, 389 ff.; *see also* Correlation, coefficient of
Correlation, multiple, 530 ff.; preliminary analysis, 533; use of multivariate estimating equation, 536 ff.; coefficient of, 543 ff.; correction for number of constants involved, 544; test of significance of the coefficient of, 544; standard error of the coefficient of, 545; application of method, 547; limitations of procedure, 563; simplification of normal equations, 652 ff.
Correlation, non-linear, 404 ff.; use of reciprocal relations, 582; use of logarithmic relations, 575 ff.
Correlation of time series, 380 ff.; of secular trends, 381; of deviations from trend, 385; dangers of procedure, 388, 389; concurrent cycles, 391; use of moving average in, 398; of short term fluctuations, 398 ff.
Correlation, partial, 584 ff.; relation to simple correlation, 549; systematic computation of coefficients of, 554 ff.; standard error of the coefficient of, 560
Correlation, rank, 374 ff., 521

Correlation ratio, 413 ff., 519; formula for, 414; computation of, 415, 417 ff.; significance of, 421; correction of, 421; relation to correlation coefficient, 422
Cost of living index, 221
Cotton statistics, 161; correlation of price and production, 384, 400
Coverage, of production index numbers, 316
Cox, G. M., 688
Criteria of curve type, 444
Crum, W. L., 302
Cumulative charts, 44 ff.; arrangement of data, 77; frequency curve, 80
Curve fitting, by least squares, 246 ff.; linear, 246; parabolic, 253, 260 ff.; of linear business series, 257; use of logs in, 264, 269
Curve type, criteria of, 444; *see also* Functional relationship
Cutts, Jesse M., 218
Cycles, correlation of, 382, 389
Cycles of reference, *see* Reference cycles
Cyclical fluctuations, correlation of, 382 ff.; *see also* Cyclical variation
Cyclical variation, 230, 302, 380, 526; removal by moving averages, 236, 284; measurement of, 293 ff.; in industrial activity, 312 ff., 390

Davenport, Donald H., 219, 254, 437, 573
Davenport, E., 415
Day, Edmund E., construction of index of physical volume, 310
Decile, graphic location of, 114
Deflation, in time series analysis, 279
Degrees of freedom, in variance analysis, 491, 496, 504 ff., 512, 517, 528; in statistical induction, 604, 611
Degree of relationship, *see* Relationship, measurement of
Deming, W. E., 469, 470; Chi-square test, 629
Dennis, Samuel J., 218
Dependent variable, *see* Variable
Depreciation, 79
Derivative, partial, 639, 643
Descartes, René, 8
Description, of frequency distributions, 86, 137 ff., 448 ff.; methods of, 99 ff.; statistical, 452 ff.
Deviate, normal, 437
Deviation, 97; probable, 152; from trend, 263, 385, 395; from mean, 347; vertical and horizontal, 363; from moving averages, 398; from means of arrays, 417; quartile, *see* Quartile deviation; mean, *see* Mean deviation; standard, *see* Standard deviation; root-mean-square, *see* Root-mean-square deviation
Differences, finite, 275
Discount rates, relation between, 340 ff., 361 ff.
Dispersion, 99, 115, 137, 414; zone of, 89 ff., 349; measures of, 137 ff., 330, 335; in correlation analysis, 490; test of differences in dispersion, 492; *see also* Variation and Scatter
Distribution, frequency, 50 ff.; description of, 137 ff.; general characteristics of, 97 ff.; of income, 71; of sawmills, 84; of heights, 87; of astronomical errors, 89; of artillery shots, 90; of coin throws, 91; of economic data, 93 ff.; of exchange rates, 95; of wage earners, 96; of bonds, 116; of stock prices, 125; *see also* List of charts
Doolittle solution, of normal equations, 540, 655
Dow-Jones index number, 393

Edgeworth, F. Y., 204; binomial expansion, 432
Elderton, W. P., Chi-square table, 624
Equation of regression, *see* Regression
Error, normal curve of, *see* Normal law of error
Error, sampling, *see* Standard error
Estimate, making of, 332 ff., 566 ff.; zone of, 349, 571 ff., 590 ff.
Exchange rates, distribution of, 94
Expected value, 294
Exponential curve, 19, 28, 258, 266, 271; modified, 272, 667; *see also* Logarithmic curve
Exponent, logarithmic, 23

Exports, statistics of, 36, 38
Extrapolation, 264, 277, 675
Ezekiel, Mordecai, 412, 477, 547, 564, 652; multiple correlation analysis, 537 ff.; correction of standard error, 542; correction of the correlation coefficient, 544

Factor reversal test, 199
Falkner, Helen D., 288
Farm price index number, 222 ff.
Federal Reserve Board, index of production, 316 ff.
Fisher, Arne, 448
Fisher, Irving, 181; time reversal test, 190; weighted index numbers, 195, 196, 204; factor reversal test, 199; ideal index number, 201
Fisher, R. A., 270, 479, 603; statistical population, 456; null hypothesis, 475; analysis of variance, 490 ff.; z table, 499, 704–5; extension of z table, 518; t table, 603, 700; significance of the correlation coefficient, 611; Chi-square table, 625, 628, 703
Frazier, Edward K., 105
Frequency curve, 41, 82; polygon, 41, 67, 85, 88, 91, 93
Frequency distribution, 50 ff.; purpose of, 56; comparison of, 86; general characteristics of, 97 ff.; *see also* Distribution
Frequency, theoretical and actual, 431 ff.
Friedman, Milton, 521
Functional relationship, 12, 389; linear, *see* Linear relationship; parabolic, *see* Parabolic relationship

Galton, Francis, lines of regression, 359
Gannt, H. L., progress chart, 46
Gauss, Karl Friedrich, normal law of error, 435
Geometric mean, 125 ff.; definition of, 125; computation of, 126; characteristics of, 127, 135; as measure of central tendency, 129; as average of relative prices, 185; of logarithmic observations, 584
Geometric progression, 18, 28, 271, 275, 669
Glover, James W., 269, 437
Gompertz curve, 272, 671
Goodness of fit, 447; criteria of, 276; Chi-square test of, 626 ff.
Graphic method, of locating averages, 120 ff.; in multiple correlation, 564
Graphic presentation, 8 ff.; of frequency distributions, 63; of time series, 227
Grouping of data, 53, 112; ungrouped data, 109; effect on mode, 115; in correlation tables, 340, 354
Growth curves, Gompertz, 272, 671; modified exponential, 272, 667 ff.; logistic, 272, 675 ff.

Hall, Lincoln W., 288
Harmonic equation, 579
Harmonic mean, 132 ff.; characteristics of, 135; of relative prices, 186; of reciprocal observations, 585, 587
Hart, Hornell, reliability of a percentage, 483
Height distribution, 87, 360
High contact, of frequency distributions, 443
Histogram, 64 ff.; *see also* Column diagram
Homogeneity, 487; tests for, 120, 630 ff.; in time series, 301; in sampling procedure, 462, 607; Chi-square test of, 633 ff.
Hotelling, Harold, 378, 479
Hyperbolic curve, 16, 28, 569

Ideal index, 201; for the measurement of production, 307
Income distribution, statistics of, 71, 97, 102
Independence, tests of, 630 ff.
Independent variable, *see* Variable
Index numbers, 18; nature of, 161 ff.; "ideal," 201; use of aggregates, 165; of retail price, 220; of cost of living, 221; of farm price, 222 ff.; of seasonal variation, 287 ff.; of industrial activity, 312 ff., 390, 393; of stock prices, 390 ff.
Index numbers of production, 305 ff.; unadjusted, 306; adjusted, 310;

Federal Reserve Board index, 316; derived from price indices, 319 ff.; of industrial productivity, 321
Index numbers of wholesale prices, 167, 216 ff.; purpose of, 170; construction of, 180 ff., 208 ff.; aggregative type, 181; arithmetic average type, 183; weighted, 196; of farm crop prices, 182, 189; geometric average type, 185, 198; median type, 185; harmonic type, 186; comparison of types, 188; time reversal test, 190; weighted types, 193 ff., 198; alternative types, 204 ff.; commodities to be included in, 209
Index of correlation, *see* Correlation index
Index of variability, 157
Induction, statistical, 452 ff., 598 ff.; nature of, 453; measures of reliability, 464 ff.; generalizing from small samples, 598 ff.
Industrial change, measurements of, 322
Inference, statistical, *see* Induction
Interaction, of principles of classification, 688
Interpolation, 70, 81, 277; for the median, 114; for the mode, 118; for monthly trend values, 273; in Fisher's z table, 500; double interpolation, 507
Irrigation, correlated with alfalfa yield, 404 ff.

Jones, D. C., binomial distribution, 660

Karsten, Karl G., 278
Kelley, Truman L., 206; reliability of constants, 485
Kendall, M. G., 629
Keynes, J. M., random sampling, 461
Killough, H. B., 569
Knibbs, Sir George, 214
Kurtosis, 100, 137, 159
Kurtz, Edwin, 77

Lag, in time series analysis, 390 ff.; changes in different cycle phases, 397
Laspeyre's index number, 193, 214
Law of large numbers, 455
Least squares, method of, 246 ff., 638 ff.; applied to linear relations, 246, 328, 354, 366, 509; applied to power curves, 260, 405; applied to logarithmic curves, 264 ff.; in correlation analysis, 366, 373, 405
Leptokurtic, 449
Life table, 77
Line of regression, *see* Regression
Linear correlation, *see* Correlation, linear
Linearity, test for, 423; by variance analysis, 508 ff.; *see also* Linear relationship
Linear relationship, 14, 16, 26, 325 ff.; fitting by least-squares, 246 ff.; in business series, 257, 268; between discount rates, 348; tests for, 423, 477, 508
Link relatives, 204
Logarithmic, equation, 26 ff., 563, 569 ff., 671; mean, 128; *see also* Geometric mean; paper, 131, 227; deviation, 265; function of the correlation coefficient, 614
Logarithms, common, 23 ff., 492, 572; use in computing the geometric mean, 125, 130; use in curve fitting, 264 ff., 269; Naperian, 435, 492; Appendix table X, 709
Logistic curve, 272, 675

Macaulay, F. R., 185, 244
Malenbaum, Wilfred, 565
Mantissa, 24
Manufactured goods, rôle in price movements, 213
Mean, arithmetic, *see* Arithmetic mean; geometric, *see* Geometric mean; harmonic, *see* Harmonic mean
Mean deviation, 139 ff.
Mean product, 351, 358
Measurement of, central tendency, *see* Central tendency; relationship, *see* Relationship, etc.
Median, definition of, 102; location of, 109 ff.; computation of, 113; graphic location of, 120 ff.; characteristics of, 134; relation to mean

deviation, 140; of relative prices, 185; standard error of, 472
Merriman, Mansfield, 90
Mesokurtic, 449
Minor, J. R., 556
Mitchell, W. C., 93, 173, 176, 212, 242, 303; comparison of index numbers, 209; business cycles, 467, 483
Mode, 96; definition of, 101; location of, 115 ff.; graphic location of, 120 ff.; characteristics of, 135
Moments, of frequency distributions, 440; about the mean, 442
Monthly trend values, 272 ff.
Mortality tables, 80
Moving average, 234 ff.; application to non-linear series, 239; measurement of seasonal fluctuations, 285 ff.; use in correlating cycles, 398 ff.
Mudgett, Bruce D., 216
Multiple correlation, *see* Correlation, multiple
Multiple frequency table, 289

Napierian logarithm, 435
National Bureau of Economic Research, 244, 320, 397; study of income distribution, 132; construction of index numbers, 219; study of production change, 309
National Industrial Conference Board, cost of living index, 221
Natural number, 24, 28; table of squares of, 706; sums of powers, 708
New York Census of Manufactures, 309, 317, 371
Non-linear correlation, *see* Correlation, non-linear
Non-linear relationship, 404 ff.; *see also* Parabolic and exponential function
Normal deviate, 437, 599; table of, 603, 699
Normal equations, for linear relationships, 249; parabolic, 254; of multivariate relationships, 537 ff.; derivation of, 639; formation of, 640; checks on, 648; Doolittle solution of, 654
Normal law of error, 98, 153, 425 ff., 435 ff.; assumptions underlying, 436; its use, 438; economic application of, 440 ff.; criteria for, 441; fitting the normal curve, 445 ff.; distribution, 332, 371, 458; departure from, 374, 378; computation of theoretical frequencies, 446; generalization of results, 448; of the distribution of means, 464; use in measures of reliability, 464 ff.; area under, 437, 699; test of goodness of fit of, 627
Null hypothesis, 475

Ogive, 80 ff., 85
Organization of data, 51, 82, 100; in time series, 226
Origin, arbitrary, 107, 351; at point of averages, 353, 365
Orthogonal polynomials, 270

Paasche's index number, formula for, 195, 215
Pabst, Margaret, 378, 479
Parabolic curve, 16, 21, 27, 270, 577; *see also* Parabolic function
Parabolic function, fitting of, 253 ff.; second degree, 260, 405; logarithmic, 264, 269, 270; testing parabolic hypothesis, 514 ff.
Parameter, 457
Pareto, Vilfredo, law of income distribution, 132
Partial correlation, *see* Correlation, partial
Peake, E. G., 94
Peakedness, 100; *see also* Kurtosis
Pearl, Raymond, 271, 272; formation of normal equations, 642; logistic curve, 675
Pearson, Karl, 156, 158, 254, 436; coefficient of correlation, 335; correlation ratio, 413 ff.; curve types, 448; descriptive measures of frequency distributions, 448 ff.; statistical inference, 454; Chi-square distribution, 618 ff., 626
Percentages, difference between and significance of, 483
Percentile, 114
Periodic fluctuation, 230; removal by moving averages, 235; *see also* Seasonal and cyclical variation

Periodic function, 21
Persons, Warren M., 204; analysis of cycle lags, 390 ff.
Platykurtic, 449
Polynomial, orthogonal, 270; *see also* Parabolic function
Population, statistical, 453, 454, 456
Potential series, 21
Power series, 253; *see also* Parabolic function
Price relative, 162; arithmetic average of, 183
Price, wholesale, 93, 168; index numbers of, 161 ff., 167, 216 ff.; *see also* Index number; price ratios, 171 ff.; measurement of change of, 174; wholesale groups, 211; index of retail, 220; of farm products, 222; deflation of, 279; measurement of variation in, 493
Probable error, 152, 155; of index numbers, 206; *see also* Standard error
Probability, 603; principles of, 425 ff.; addition of, 427; measurement of, 429; a priori, 431; empirical, 431; normal, 439, 459, 471; normal table of, 699; integral, 436
Probability, curve 98; *see also* Normal law of error
Probable value, 332
Production, statistics of, 10, 35, 40, 43, 47; of fuel, 163, 265; of crops, 192; as measured by index numbers, 305 ff.; *see also* List of charts
Product-moment method, 349 ff., 368; for classified data, 354 ff.
Projection, of trend values, 277, 402
Purposive selection, in sampling procedure, 462

Quartile, 114; graphic location of, 120 ff.; deviation, 150 ff., 154; standard error of, 473

Random fluctuations, 231; removal by moving averages, 241
Random sampling, 458, 461; *see also* Sampling
Range, of variation, 139, 154; semi-interquartile, 151
Rank correlation, 374 ff.; *see also* Correlation, rank
Rate, of interest, 30, 76, 228; of change, 40, 267, 278, 587; of exchange, 94; averaging of, 125
Ratio, chart, 29, 35
Ratio, correlation, 413; *see also* Correlation ratio
Reciprocals, use in measuring relationship, 578 ff., 675 ff.
Reed, Lowell, J., 272; logistic curve, 675
Reference cycles, 243, 262; correlation of, 382
Regimen, 214, 322
Regression, lines of, 359 ff.; use of, 364 ff., 367, 423, 607; coefficient of regression, 359 ff., 363, 479, 561, 607; for cotton production and price, 387, 401; standard error of coefficient of regression, 479, 607, 609
Relationship, between income and auto registration, 326 ff., 352; measurement of, 325 ff., 334; between discount rates, 340 ff.; between time series, 380 ff.; temporal, 391 ff.; linear, *see* Linear relationship
Relative deviations, 129; weighted, 167
Relative price, 162; arithmetic average of, 183; geometric average, 185, 198; harmonic average, 187; weighted average, 196
Relative variation, measurement of, 156 ff., 264
Reliability, measures of, 464; of the mean, 464; of the difference between means, 481, 483; of the median, 472; of the standard deviation, 473; of the coefficient of correlation, 474; index of correlation, 477; coefficient of regression, 478
Residuals, 247
Residual variability, *see* Variability, residual
Retail price, index of, 220
Richardson, A. H., 49
Rietz, H. L., 143
Robertson, R. D., 405
Robinson, G., 88, 274, 465
Root-mean-square deviation, 146,

276, 330, 416; *see also* Standard deviation
Rulon, P. J., 485

Sample, size of, 117; estimates from, 146; in constructing index numbers, 206
Sampling, problem of, 452 ff., 460; random, 458, 461; generalizing from small samples, 598 ff.; errors of, 293, 447; *see also* Standard error
Sasuly, Max, 270
Scale, for curve reading, 39
Scatter, 99; degree of, 137, 334, 409, 414, 646; *see also* Variation
Scatter diagram, 326, 328, 348, 370, 416
Scott, Frances V., 219
Seasonal variation, 230, 284 ff., 380; removal by moving averages, 235; measurement of, 287 ff.; adjustment of, 317; test of significance of, 522 ff.
Secular trend, 229, 380, 487; of cotton production and price, 383, 385; measurement of, 231 ff.; representation by moving average, 234; by mathematical curves, 244 ff., 667 ff.; of business series, 257 ff.; selection of curve, 274 ff.
Selection of curve of trend, 274
Semi-interquartile range, 151
Semi-logarithmic charts, 28, 264; advantages of, 40
Series, periodic, 21; potential, 21; continuous, 75
Sheppard, W. F., correction for grouping, 150, 442 ff.; table of normal areas, 436
Shewhart, W. A., 49; distribution of the standard deviation, 600 ff.
Significance, tests of, 464 ff.; *see also* Standard error
Significant figures, 485
Sine curve, 21
Skewness, 96; measures of, 100, 122, 137 ff., 157 ff., 449; of geometric series, 129; of the standard deviation, 600; of the correlation coefficient, 610
Slope, 293; of regression line, 336, 350, 359, 361; *see also* Regression coefficient
Smoothing of curves, 69 ff., 76, 117
Snedecor, George W., 449, 688
Snyder, Carl, 229
Spurr, W. A., 293
Squares of natural numbers, table of, 706
Standard deviation, 145 ff., 330, 371, 416; characteristic features of, 155; use in adjusting index numbers, 311, 393, 395; in terms of moments, 443; about the means of arrays, 418; use of, in variance analysis, 491, 494; *see also* Standard error
Standard error, of the binomial distribution, 434, 660; of the mean, 464, 664; of the difference of means, 481, 483; of the median, 472; of the standard deviation, 473; of the correlation coefficient, 474, 545; of the correlation index, 477; of the regression coefficient, 478; of the partial correlation coefficient, 560, 615; of the z function, 493, 615; limitations of above measures, 486 ff.
Standard error of estimate, 330 ff.; computation of, 333, 338, 370, 388, 401, 406, 590; short-cut calculation, 346, 354; of parabolic functions, 410; significance of, 349, 371; about line of regression, 480; correction of, 413, 542; in multiple correlation analysis, 534, 541 ff.; of logarithmic functions, 571 ff.; in ratio terms, 573; in reciprocal terms, 581; zones of estimate, 590 ff.
Starr, G. W., 312
Statistic, 457
Statistical description, *see* Description
Statistical induction; *see* Induction
Steinmetz, C. P., 256
Stewart, Ethelbert, 84
Stock price cycles, relation to business activity, 390, 397
Straight line, fitting of, 246; *see also* Linear relationship
Stratification, in sampling procedure, 462
Stratified purposive sampling, 463; standard error of, 472

"Student," standard error of the rank correlation coefficient, 479; standard error of the mean, 599; distribution of the standard deviation, 600 ff.
Sturges, H. A., 57
Symbols, glossary of, 691 ff.
Symmetry, 100; degree of, 120; *see also* Skewness

Table, of areas under the normal curve, 699; Fisher *t* table, 603, 700; of significant values of the correlation coefficient, 612, 701; of relations of the correlation coefficient to the *z* function, 702; of the distribution of *z*, 704–5; of the powers of natural numbers, 706, 708; of common logs, 709
Tabulation of data, 51, 62; in correlation tables, 341, 354, 415
Tendency, central; *see* Averages and Central tendency
Thompson, F. L., 292
Time reversal test, 190
Time series, charts, 33, 48, 50; analysis of, 225 ff., 295; graphic representation, 227; removal of cycles, 234; fitting a line to, 252; measurement of seasonal fluctuation, 284 ff.; of cyclical fluctuation, 284; measurement of relations between, 380 ff.; *see also* Correlation of time series
Tolley, H. R., 537, 652
Trend, 262; of price movements, 170; of monthly values, 272; selection of curve of, 274 ff.; measurement of, 225 ff.; secular, *see* Secular trend

Ungrouped data, 109; product moment method for, 352
Uniformity of nature, principle of, 457
Unweighted index number, 184
U. S. Bureau of Internal Revenue, 326
U. S. Bureau of Labor Statistics, statistics of fuel production, 164; index of wholesale prices, 168, 172, 176, 212, 216 ff., 282; index number used, 193; index of retail prices, 220; cost of living index, 221
U. S. Department of Agriculture, index of farm prices, 222

Variability, measures of, 490 ff., 560; between classes, 494 ff.; absolute, 586; residual, 526, 689; *see also* Variance and variation
Variable, 11; relations between variables, 325 ff., 359, 360
Variance, analysis of, 490 ff.; *z* test of difference in variability, 492, 506, 513, 517; in testing variability between classes, 494; in the measurement of relationship, 501 ff., 519; in testing linearity, 508 ff.; curvilinear hypothesis, 514 ff.; testing seasonal fluctuation, 522 ff.; in testing the multiple correlation coefficient, 545; in testing significance of principles of classification, 681 ff.
Variation, 97; measures of, 99, 137 ff., 330; absolute, 138; comparison of measures of, 153, 155; measures of difference in, 490 ff.; coefficient of, 156; in price relatives, 171 ff.; within and between arrays, 502; *see also* Seasonal and cyclical fluctuation
Verhulst, P. F., 272

Wage statistics of, 96, 103, 105, 111, 124
Wahr, George, 437
Walsh, C. M., 130, 201; ratio variability, 596
Weighted average, 104, 106; of relative prices, 106; geometric, 125; moving average, 244
Weldon, W. F. R., dice experiment, 432, 618
Wheat, exports of, 33; yield correlated with fertilizer, 415
Whipple, G. C., 87
Whittaker, E. T., 88, 274, 465
Wholesale price, 211 ff.; index of, 216 ff.; *see also* Price
Working, Holbrook, harmonic mean, 588

Yates, F., 688
Yule, G. U., 60, 418; Chi-square frequencies, 622, 629

Zone, of estimate; *see* Estimate, Dispersion
z test of variability, 492, 506, 514, 517; tables of, 704–5; standard error of, 615
z transformation of correlation coefficient, 613 ff., 702