UNIVERSAL
LIBRARY

OU_148623

UNIVERSAL
LIBRARY

INTRODUCTION
TO
ECONOMIC STATISTICS

This book is produced in full compliance with the government's regulations for conserving paper and other essential materials.

INTRODUCTION

TO

ECONOMIC STATISTICS

BY

WILLIAM LEONARD CRUM

Professor of Economics, Harvard University

ALSON CURRIE PATTON

Associate Actuary, The Massachusetts Protective Life Assurance Company

AND

ARTHUR ROTHWELL TEBBUTT

Professor of Business Statistics, Northwestern University

FIRST EDITION

FIFTH IMPRESSION

McGRAW-HILL BOOK COMPANY, INC.

NEW YORK AND LONDON

1938

COPYRIGHT, 1938, BY THE
McGRAW-HILL BOOK COMPANY, INC

PRINTED IN THE UNITED STATES OF AMERICA

All rights reserved. This book, or parts thereof, may not be reproduced in any form without permission of the publishers.

THE MAPLE PRESS COMPANY, YORK, PA.

PREFACE

This text is a complete revision of a book bearing a similar title which was published in 1925 by the A. W. Shaw Company and subsequently by the McGraw-Hill Book Company, Inc. The revision has included many changes suggested by helpful critics of the former edition and has aimed at bringing the illustrative material up to date and extending it to a somewhat wider range of economic topics. Moreover, certain portions of the present edition—in particular the first three chapters—are almost entirely new. In certain parts of the treatment, materials which appear now appropriate for more advanced texts have been eliminated or reduced in emphasis. The intent throughout has been to preserve the elementary character of the treatment.

The book has been prepared while keeping constantly in mind the needs of those students who are interested in the application of statistical methods to economic problems. This point of view has controlled the selection and arrangement of the topics covered and is responsible for the almost exclusive use of economic statistics for illustrative material. No attempt has been made to discuss numerous recent developments in statistical technique, because they are not regarded as coming within an elementary treatment which aims at expounding fundamentals.

The presentation is designed to enable the student who does not have extensive mathematical training to become familiar with statistical processes and acquire a working knowledge of the limitations in the use of such processes in economic analyses. The reader of the book is assumed to have more than a reminiscence of the ground covered in elementary algebra, and he should also have or shortly acquire some knowledge of logarithms and the slide rule. Further mathematical training is unquestionably desirable for one who hopes to become a statistical specialist; but as nearly all symbolic developments of the book are confined to footnotes, a good understanding of the subject can be secured with no mathematical background beyond the minimum mentioned above.

When the book is used as a text in an organized course, the material of the entire text (except the final chapters in Part II

and in Part III) can be covered in one half year. This presupposes at least three meetings each week, of which one is set aside for exercise work in the "statistical laboratory." To carry out such a program, it will be necessary to cover the chapters of Part I, which is less difficult than the other parts, rather rapidly. If the book is to be used for a year course, the work of the text will need to be supplemented by references to the more elaborate treatments of the advanced topics in other sources or by the assignment of fairly complicated report topics.

We do not believe that anyone, whether a member of a class or a self-taught student, can learn statistics by reading a textbook. Actual performance of the routine details involved in statistical analyses is essential; every student should do a generous number of exercises and early form the habit of completing each task rather than assuming that he understands the unfinished portion.

Every care has been taken to safeguard the quoted data and the computations and charts from error, as well as to ensure textual accuracy. It is not taken for granted, however, that the result is perfect or even as nearly perfect as might reasonably be desired, and the authors will therefore be grateful for information of any errors which are noticed and for all suggestions for the improvement of form or content.

In the work of revision, Mr. Crum has assumed main responsibility for Parts I and II and Mr. Tebbutt for Part III; Mr. Patton has given attention to the entire manuscript. Mr. Frank F. Dodge, of the staff of the Paul Revere Life Insurance Company, Miss Dorothy Wescott, assistant editor of the *Review of Economic Statistics*, and Mrs. Harriet S. Ross, formerly of the staff of that *Review*, have given painstaking and discerning assistance in assembling and criticizing the statistical materials. The authors are grateful also for generous and painstaking aid rendered by Mrs. Esther S. Morton and Miss Eleanor Lyons in preparing the charts, and by Miss Althea MacDonald, Mrs. Anna H. Thorpe, and Mrs. Dorothy Miller in preparing drafts of text and tables.

W. L. C.
A. C. P.
A. R. T.

CAMBRIDGE, MASS.,
WORCESTER, MASS.,
PROVIDENCE, R. I.,
August, 1938.

CONTENTS

PART I. STATISTICAL DATA

PAGE

PREFACE v

I

INTRODUCTION

The Task of the Economic Statistician 3
Statistical Items 5
Statistical Series 11

II

VARIABLES AND HOMOGENEITY

Economic Variables 18
Homogeneity 26

III

SOURCES AND THEIR USE

Primary and Secondary Data 38
Primary and Secondary Sources 39
Quality of Sources 45
Compilation from Several Sources 48

IV

COLLECTION OF PRIMARY DATA

Planning a Statistical Investigation 50
The Primary Survey and the Informants It Reaches 51
Methods of Conducting the Survey 53
The Blank Form 55
Editing the Returns 57

V

CONSTRUCTION OF GENERAL TABLES

General and Summary Tables 60
The Classification Scheme in a General Table 61
Orders of Tabulation 66
Practical Details in Table Construction 70

VI

SUMMARY TABLES

Working and Publication Tables 75
Preparation of Working Tables 76
Preparation of Publication Tables 81

PAGE

CHARTING: CATEGORICAL SERIES

Three Methods of Summarizing Data 91
Limitations of Graphic Presentation 92
Kinds of Charts for Categorical Data 93

VIII

CHARTING: TIME SERIES

Use of Charts in Analysis and Exposition 102
The Location of Points with Arithmetic Scales 102
The Construction of a Grid 104
Actual and Percentage Variation 108
The Ratio Scale 112
Interpolation and Smoothing 114
Comparison of Time Graphs 116

IX

CHARTING: FREQUENCY SERIES

The Frequency Polygon and Its Main Features 119
The Block Diagram 128
The Statistical Sample, and Smoothing 130
Ogives 135
Discrete Frequency Series 139
Weighted Frequencies 144

PART II. GENERAL ANALYTICAL METHODS

X

SUMMARY NUMBERS

General Properties of Averages 155
Types of Averages 159
Rates and Ratios 166

XI

AVERAGES FOR FREQUENCY SERIES

Averages of Discrete Frequency Series 168
The Mean of a Grouped Frequency Series 174
The Skeleton Method of Computation 177
Properties of the Mean 181
The Median of a Grouped Frequency Series 182
Properties of the Median 184
The Mode of a Grouped Frequency Series 185
Other Averages of Grouped Frequency Series 190

XII

DISPERSION

Measures of Dispersion Derived Directly from the Array 193
The Average Deviation 200

PAGE
The Standard Deviation 201
Coefficients of Dispersion 206

XIII

THE NORMAL LAW OF ERROR

The Equation and Properties of the Normal Curve. 208
Fitting a Normal Curve to a Given Series 209
Standard Units 215
Measures of Dispersion, and the Probable Error. 215
Selection of a Random Sample. 216
Probable Errors of the Characteristics 218

XIV

SKEWNESS, MOMENTS

Coefficients of Skewness 221
Extremely Skew Series 223
Moments of a Frequency Series 227
Curve Fitting by the Use of Moments 228

XV

THE CORRELATION TABLE

Definition of Correlation 230
Description of the Correlation Table and Scatter Diagram . . . 231
The Lines of Regression 235
The Distribution of Frequency within Each Array 237

XVI

THE COEFFICIENT OF CORRELATION

The Correlation Coefficient as Based upon the Scatter Diagram . . 239
Calculation of r from the Categorical Table. 241
Derivation of r from a Correlation Table 244
The Skeleton Method 245
The Line of Regression 247
The Errors of Estimate 249

XVII

FURTHER CORRELATION METHODS

The Correlation Ratio 253
Spurious Correlation 254
Partial Correlation 255

PART III. THE ANALYSIS OF TIME SERIES

XVIII

RELATIVES AND INDEX NUMBERS

Commodity Prices 263
Price Relatives 264
Graphic Representation of Price Changes. 267

PAGE
Index Numbers as Averages of Price Relatives . . . 273
Simple Means of Prices or Price Relatives . . . 277
Other Simple Averages . . . 280
Fixed-base and Chain Indexes . . . 282

XIX

WEIGHTED INDEX NUMBERS

The Necessity of Weighting . . . 283
The Purpose of an Index as a Determinant of the Weighting System . . . 284
Values in Exchange as Weights . . . 285
The Calculation of Weighted Indexes . . . 287
Type Bias . . . 290
Weight Bias; the Joint Effect of This and Type Bias . . . 291
Professor Fisher's Ideal Index . . . 292
Further Tests of Indexes . . . 293
Constant Weights More Practical Than Variable Weights . . . 294
Class and Commodity Weights . . . 295
Statistical Deflation . . . 296

XX

SECULAR TREND

Four Components of Time Fluctuation . . . 298
Graphic Examination of the Trend . . . 300
Determination of the Line of Trend by Computation . . . 305
The Ordinates of Trend . . . 311
The Elimination of Trend . . . 316
Negative, Broken, and Curvilinear Trends . . . 317

XXI

SEASONAL VARIATION

Graphic Evidence of Seasonal Tendency . . . 326
The Distribution of Link Relatives . . . 330
The Median-link-relatives . . . 335
The Determination of Seasonal Indexes . . . 338
The Elimination of Seasonal Variation . . . 342
Analysis of Quarterly Data . . . 345

XXII

CYCLICAL FLUCTUATIONS

Cyclical Fluctuations in Different Series . . . 353
Differences in Form and Extent of Cyclical Swings . . . 356
Cycles Expressed in Standard Units . . . 358
The Business Cycle . . . 361

XXIII

LAGGING CORRELATION AND FORECASTING SEQUENCE

Correlation When No Lag Exists . . . 363
Numerical Determination of Lag . . . 365

PAGE
The Forecasting Sequence 367
The Nature of the Sequence 368

APPENDIX

A. NUMERICAL DATA FOR CERTAIN OF THE CHARTS OF PARTS I TO III 371

B. LABORATORY PROCEDURE
Transcription 391
Computation 400
Charting 405

C. THE NORMAL CURVE OF ERROR, AND OTHER FREQUENCY CURVES
The Parameter of the Curve 407
The Quartiles of the Normal Curve 408
The Average Deviation for the Normal Curve 409
Distributions Which Are Nearly Normal 410

D. LOGARITHMS OF NUMBERS 413

INDEX 417

PART I

STATISTICAL DATA

CHAPTER I

INTRODUCTION

THE TASK OF THE ECONOMIC STATISTICIAN

The economist is necessarily interested in statistical data and statistical methods because they are indispensable aids in his understanding of economic problems. Economic science is concerned with the production and distribution of wealth and with all the characteristics and peculiarities of the complicated human and physical organization by which wealth is made available for consumption and for use in further production of wealth. The problems of the economist unavoidably involve *numerical* magnitudes—such as output of wheat, iron, clothing, volume of exports, wages, prices, profits, bank deposits—which are subject to *variation*, from time to time or between places or among particular cases. Moreover, these numerical magnitudes and their variations are subject to interrelations which are more or less definite and more or less complex. Accordingly, the economist's study includes not only the variations in a particular magnitude but also the way in which and the degree to which the variations in that magnitude are related to variations in other magnitudes.

Such are precisely the problems requiring facts in the form of statistics and analysis of facts by statistical methods. We must not suppose that statistics are the sole factual aid of the economist, or that statistical data and their analysis will inevitably yield a complete and conclusive solution of all economic problems. But we may confidently rely upon statistics as an aid in studying many economic problems, and we may even expect to secure from statistics a tolerably final answer to a considerable number of economic questions. The first task of the economic statistician is to give numerical definiteness to economic concepts and economic problems, by stating them in terms of statistical data which reflect the actual economic conditions and developments of real life. But his task does not end at that point: it includes also the development and use of a technique and theory by which he can analyze and summarize the stated statistical facts, draw inferences from them, and indicate their economic implications. To accomplish

this broad task, he must acquire an extensive knowledge of the sources and limitations of statistical information. He must secure a firm grasp on the arithmetical operations appropriate for analyzing and summarizing such information and on the theoretical validity of these operations. And he must also know economic theory and its applications so well that he can bring his statistical results and his economics together into an effectively integrated body of knowledge.

Although the third of these "must" points will receive little specific stress in this book, it has been kept constantly in mind in preparing the book; and the student is warned that a mere study of statistical data and statistical techniques, apart from the setting of the science—in this case economics—to which they are to be applied, is a dangerous intellectual adventure. This is not to say that statistical science cannot exist by itself, or that many of the techniques discussed below cannot be applied effectively in other sciences than economics. The essential point is that, for the economist, statistics must be studied and understood in the light of economic problems and of the peculiarities and complexities which affect the variation of numerical magnitudes in economics. For the economist, statistics must be regarded as a tool, intelligently fitted to the use he needs to make of it.

In approaching his first task, that of mere statistical description of economic concepts and economic problems, the economic statistician encounters many intricate difficulties which require a careful application of specific techniques. It is a mistake to regard the mere collection and organization of statistical facts, for the numerical description of economic situations or developments, as a routine task within the capacity of an untrained person. The blind assembling of statistical data, and the superficial taking for granted that such data are pertinent to the particular economic question in hand, must at best yield a statistical picture which is not surely conclusive, and at worst provide the economist with so-called facts which are positively misleading. Only by the most painstaking criticism of the sources from which the statistics are obtained, of the precise significance of the definitions which specify and describe the numerical items, and of the comparability and consistency of the items, can a body of statistics be rendered trustworthy for economic interpretation. Substantially the same rules of evidence apply to the use of statistical facts as to that of any other facts. But the need of careful testing is especially great in the case of statistics because of the unfortunate tendency of the

human mind to attribute to a mere numerical figure or table or chart a peculiar finality and accuracy—to assume that it is in a peculiar degree a "cold fact." These remarks are made, at somewhat tiresome length, because the sober truth is that much of the faulty reasoning and many of the nonsensical conclusions which economists base upon statistics are traceable to imperfections in the statistics themselves.

Faced with a particular economic question, the economist must resist a strong temptation to seize the most readily accessible source of statistical information, copy off a table or chart which seems to bear an appropriate title, or grasp some statistical figure which wears a name suggesting its pertinence. Such a superficial practice might, and probably would, lead him into the same dangers of drawing erroneous inferences as do in fact mark much of the so-called statistical work currently done in economics. Such a practice explains why so many statistical arguments in economics are bad: they are bad because the economist is using partially or wholly bad statistics and is through his own negligence unaware that he is doing so. It is generally not a matter of statistical *analysis*—of using specially defined statistical summarizations such as dispersions, correlation coefficients, or index numbers—but a mere matter of statistical *data*. The beginner in economic statistics therefore makes a serious mistake if he devotes attention to questions of analysis before first securing a thorough acquaintance with statistics, the raw material of analysis. Accordingly, considerable emphasis will be placed, in the immediately following pages, upon the nature of statistical data—their sources, definitions, units, consistency, comparability, and manner of presentation. And the student is urged to give continuing thought to these points, even after he passes on to the less humdrum topics which make up the central body of statistical method, lest he later fall into the careless habit of applying refined and precise analytical methods to raw data which have no meaning or a meaning quite different from what he supposes. All the foregoing is said without meaning to minimize the grave danger of faulty reasoning about statistical facts, even facts of the best quality.

STATISTICAL ITEMS

As some of the foregoing comments have already suggested, the word *statistics* is used in two important senses: to cover statistical *data*—numerical facts—and to cover statistical *analysis*—the

theory and technique of treating statistical data. Statistical analysis includes all the technical operations incident to compiling, presenting, summarizing, discussing, and interpreting statistical data; and it also includes the theoretical foundations of such operations. Certain refinements upon this general statement of the significance of the term *statistics*, and a more specific indication of the scope and limitations of the subject will become evident in the following chapters.[1]

Statistical data consist of numerical facts. A single numerical fact may be called a statistical *item*. Occasion may require the use of an isolated statistical item, and certain of the comments below indicate the care which should be exercised in such use. But ordinarily the problem or question in hand will require the use of several, perhaps many, statistical items; and the very word *statistics*, by implication or through custom in its use, suggests a multiplicity of numerical facts—a group of items. A list or group of statistical items is called a *statistical series*. Strictly, the term *series* should not be used for an indiscriminate and disconnected list of items, but should be reserved for a list the elements of which have some connection with each other. The manner and degree of such connection may be vague and uncertain; but some connection, direct or indirect, is essential if we are to call a batch of items a *statistical series*.

These points will be clarified by the following illustrations. Each of the five numerical facts below is a statistical item.

a. Shipments of rolled and finished steel products, of U. S. Steel Corporation, 1936	10,784,273	tons
b. Silver certificates in circulation, outside U. S. Treasury and Federal Reserve Banks, end of January, 1938.....	1,085	million dollars
c. Closing price of no. 2 yellow corn, at New York, March 11, 1938.....	71¾	cents per bushel
d. Total deposits of insured commercial banks in U. S., May 13, 1937.....	45,188	million dollars
e. Cash dividends paid by U. S. corporations reporting for income tax, 1934.....	1,036,781	thousand dollars

These five items do not constitute a statistical series; except for the fact that each pertains to a particular aspect of a single economic community, there is no connection between the items. Each is an isolated numerical fact.

---

[1] For summary comment on the origin and history of the word *statistics*, and for points on the definition and scope of the subject, see G. U. YULE and M. G. KENDALL, "An Introduction to the Theory of Statistics," Chas. Griffin & Co., Ltd., London, 1937, Introduction; and A. L. Bowley, "Elements of Statistics," 5th ed., P. S. King & Son, London, 1926, Chap. 1.

Certain helpful comments can be made about these isolated facts, as separate *statistical items.*[1] Each item has attached to it a label, or title, and a unit. The label is a verbal descriptive term which tells more or less adequately the nature and significance of the item. This descriptive label is an essential part of, or attachment to, a statistical item: the mere number, by itself, is obviously useless. Moreover, the label needs to be sufficiently long and detailed to leave no essential qualification subject to guesswork: if the label is unduly abbreviated, it fails to describe the item adequately. For example, if the label of Item *a* had read "Shipments of steel products," we should not have known three important descriptive facts: the kind of steel products, the identity of the producer, and the time interval during which shipments occurred. But the titles as given do not necessarily provide a wholly adequate description of each statistical item. In the case of Item *a*, for example, we might wish to know whether the figure covers all the subsidiaries of the United States Steel Corporation and whether it includes shipments from one subsidiary to another or only shipments to outsiders. Also, the very meaning of the word *shipments* may be in doubt: Does a shipment occur when the product leaves the factory? If a product is made up of several units, or parts, does a shipment occur when each unit leaves the factory or only when the last of the units has been shipped? If the corporation has a subsidiary engaged in erecting structural steel at the site of the purchaser, does a shipment occur when the steel is sent from the factory to the erector or when the erector has completed installation? These questions suggest the type of uncertainty which may affect labels appearing from a superficial inspection to be wholly precise.

Specification of an item.—The foregoing comments bear upon a fundamental point, the necessity for a full and precise specification —descriptive definition—of a statistical item. Most of the items listed above appear to be carefully specified, but others besides Item *a* reveal, upon inspection, inadequacies in their labels. For example, in Item *c*, what is a "closing price"? (Does it relate to an actual sale or is it a "bid" or an "asked" quotation?) What is "No. 2 yellow corn"? (Can we know the significance of this term without inquiring as to quality standards of corn and the

[1] Throughout this text it should be borne in mind that the word "item" may have reference simply to the numerical part of the *statistical item* or it may be used to denote the entire statistical item. The double use of the term cannot well be avoided and in general the context is sufficiently clear to preclude any confusion.

practice of grading?) What is the significance of "at New York"? (Does it mean for corn delivered in New York and if so where and in what manner, or does it mean that the prices are made at a particular organized exchange in New York?) In the case of Item *e*, what does "reporting" imply? (We should have to find out that it means a limitation on the coverage of the item: only those corporations in the United States are covered which were required by law to file an income tax return and in fact did so.) And does "cash dividends paid" really mean that the given item is a true figure for actual dividends paid? (Examination of the source of the item shows that the published figure is compiled from tax returns before they are audited by the Treasury and subjected to correction of any errors revealed in the dividends, or other accounting information, as reported on the returns.) And how precise is "1934"? (Actually, investigation shows that it does not mean the calendar year 1934 exclusively, because many corporations file tax returns pertaining to "fiscal years" ending anywhere between July 1, 1934, and June 30, 1935, and all these would be included in the 1934 figure.)

These illustrations aim to bring out the insidious danger of taking for granted what the words of a label seem to imply. No amount of care can in all cases ensure the inclusion in the label of every point needed for a completely accurate description of a statistical item, and practical considerations necessitate drawing the line somewhere short of perfection. The time and energy needed for complete specification and the mere space required for listing all the details stand in the way of absolute precision. But this practical rule must not be made an excuse for neglecting to determine and list essential details in the specification of every statistical item. Even at the cost of painstaking research and inquiry, and of liberal use of space for writing out labels in detail, the statistician must protect himself against vague definitions. Otherwise the statistical raw materials which he uses, though they may appear as exact numbers, are in fact numbers attached to ideas which are far from exact. This task of definition is the first responsibility in every statistical job, no matter how simple or how complex.

Units.—In addition to bearing a descriptive label, each of the five items above is qualified by a statement of unit. This also is an essential element in a statistical item: a numerical figure is of no usefulness unless the unit in which it is stated is definitely known,

In some cases, convenience dictates that the unit be stated as part of the label. Thus, Item *a* might have read:

Tons of rolled and finished products shipped by U. S. Steel Corporation in 1936 10,784,273

In rare instances, the unit is unambiguously implied by the label and does not need be stated explicitly. But the general rule requires the unit to be stated specifically.

Frequently the unit is more complicated than those for Items *a* to *e*, and in some instances the description of the unit may be nearly as involved as the label of the item. The most complicated unit among the five listed above is that of Item *c*. But consider the unit for the following item:

f. Freight traffic density on the Pennsylvania Railroad in 1936 . . . . . 3,590,844 ton-miles per mile of road

Here there are three elements in the unit: the ton as a measure of the volume of traffic, the mile as an indicator of the distance freight is hauled, and the mile of road as a factor to account for the size or capacity of the railroad line. Even more intricate descriptions of units may be needed in particular cases, and in all cases the description of units must be in sufficient detail to preclude any misapprehension.

With respect to units, as well as labels, descriptions which superficially appear to be adequate may in fact be defective. For example, in Item *a*, the term *tons* is only superficially precise; as a matter of fact, we remain uncertain what kind of ton is intended—whether the "short" ton of 2,000 pounds, the "long" ton of 2,240 pounds, or some other weight called a *ton*. Unless such questions are definitely cleared away by a descriptive phrase qualifying the unit as stated, the critical user of the statistical item must obtain the unstated information. He may be able to find such information in the source from which the item was quoted, or he may trust to an inference based upon the "custom of the trade" according to which (he may know) steel is understood as measured in a ton of 2,240 pounds, or he may have to address an inquiry to some informed person.

Similarly, in Item *c* no information is presented as to the size of the measure called a *bushel;* and yet the critical statistician knows that bushels of different sizes or weights exist and are used for various purposes. The customary meaning of a bushel for

measuring grain may be so well established that the lack of a qualifying phrase in the unit of Item *c* can be ignored; but if the statistician wishes to restate the price in cents per pound he is at once obliged to find an answer to this question. In general, every unit in terms of which a statistical item is stated needs to be examined to make certain that it is adequately and unambiguously described.

Source reference.—The items listed above are seriously defective in one respect: each should be accompanied by a reference indicating the source from which it was obtained. The chief reasons for this requirement are that such a reference enables the user of the item: to verify its accuracy, to seek the most probable place for supplying information—such as more detailed specification of the item or more precise description of the unit—which is not given, to find other items comparable with or related to the given item, to form an opinion as to the trustworthiness of the item in the light of the reliability of the source. So important are the aids thus afforded that a statistical item not accompanied by a source reference should not be accepted for use.

The source references for the items above are as follows:

a. *35th Annual Report.* U. S. Steel Corporation, Year ended December 31, 1936, p. 4.
b. *Federal Reserve Bulletin*, March, 1938, p. 205.
c. *Commercial and Financial Chronicle*, March 12, 1938, p. 1746.
d. *Annual Report*, 1936, Washington, Federal Deposit Insurance Corporation, 1937, p. 71.
e. *Statistics of Income*, part 2, 1934, Washington, U S. Treasury Department, 1937, p. 56.
f. "Mundy's Earning Power of Railroads," J. H. Oliphant & Co , New York, 1937, p. 163.

and Item *a* can now be restated as:

Shipments of rolled and finished steel products, of U. S. Steel Corporation 1936*. . . . . 10,784,273 tons

The significant characteristics of sources of statistical data need careful discussion before the limitations upon data as quoted from sources can be understood; and extensive attention is given to this subject below (Chap. III). For the present, the essential requirement is emphasized that every statistical item must be accompanied by an adequate source reference. As will appear presently, this requirement can be met practically in cases involving lists or groups of items without a separate reference for each item and the attendant lavish expenditure of space and effort.

* *35th Annual Report*, U. S. Steel Corporation, Year ended December 31, 1936.

STATISTICAL SERIES

Although an isolated statistical item, or a few such items, may be adequate factual material for certain exceedingly simple economic questions, the great bulk of statistical applications rest upon groups or lists of items having some connection or basis of comparability with each other. Such a group, or list, is called a *statistical series*. For example, the statistics in the following textual statement constitute a very short and simple statistical series:

> Taking 1937 as a whole, the growth in commercial, industrial, and agricultural loans . . . amounted to $285,000,000 at central reserve city banks in New York, $400,000,000 at other reserve city banks (including Chicago), and $265,000,000 at country banks.[1]

Although, as in the above case, the items of a statistical series—if the series is very short and simple—may satisfactorily be scattered through a textual statement, a far more convenient form of presentation, and a form which is almost unavoidable if the series is very long or complicated, is the tabular arrangement.

Tabular presentation.—The data of the foregoing illustrative series might have been presented in the form of Table 1. Customarily an isolated statistical item, or several such items not so related to each other as to constitute a statistical series, would be presented in a running text; but a *series* of items is much more available for examination and analysis if it is tabulated. Even in those instances, like the above, in which a short series has in fact been presented in running text, the first step the statistician takes is to organize the items into a table such as Table 1.

TABLE 1

INCREASE IN COMMERCIAL, INDUSTRIAL, AND AGRICULTURAL LOANS AT MEMBER BANKS OF THE FEDERAL RESERVE SYSTEM DURING 1937*

Central reserve city banks in New York	285
Other reserve city banks (including Chicago)	400
Country banks	265
Total, all member banks	950

* Unit: million dollars. Source: *Federal Reserve Bulletin*, March, 1938, p. 184.

Various and important technical questions concerning the construction of statistical tables and their interpretation will be considered below (Chap. V). At this point, however, certain

[1] *Federal Reserve Bulletin*, March, 1938, p. 184.

significant facts should be noted. In organizing a table we display the data in a convenient and readily understandable form. The table includes all the essential elements noted above as pertaining to every statistical item: the numerical figure itself, the label as a specification of the item, the unit, and the source. These elements are indispensable for the items of a statistical series, whether in tabular or other form, just as surely as they are indispensable for an isolated item. But one of the advantages of the tabular form—one of the economies which it achieves—consists in covering various items under a single description or partial description of the nature of the item, the unit, and the source. Thus the labels of the three top items in Table 1 have been broken up: that portion of the label common to every item is stated once for all in the title of the table, and the portion of the label peculiarly attaching to any one item is stated opposite that item in the left margin. Likewise, the source reference and the unit designation, being common to all items, are stated once for all in a footnote to the title. By exception, the note specifying unit is sometimes given just beneath the title. Therefore the tabular form, in addition to arranging the numerical figures in an order which facilitates their comparison, economizes in the presentation of essential facts as to specification, units, and sources. For the tabular presentation of more complicated statistical series, as we shall see (Chap. VI), this facilitation of comparison and this economy in description are not completely realized; but they are always greatly aided by the tabular method. In fact, tabulation is the standard and fundamental method of presenting statistical series.

Types of series.—Tables 2 to 6 give further examples of simple statistical series. In each of these tables, as in Table 1, that portion of the label which pertains to a particular item appears in the left section, adjacent to that item. These particular designations are called *stubs*. The stubs in any one table, such as Table 2, serve to distinguish the several items from each other, and the stubs therefore constitute a *rule of classification*. The rule of classification thus appears as an essential part of a statistical series.

A *statistical series* may, then, be more specifically defined as a group of statistical items related to each other by some rule of classification. The rule may classify the items according to any significant quality or characteristic. Thus, classification is according to location of the banks in Table 1, type of product in Table 2, date in Tables 3 and 4, and size in Tables 5 and 6.

TABLE 2

EXPORTS OF HEAVY IRON AND STEEL PRODUCTS FROM THE UNITED STATES IN 1935*

Scrap	2,104.0
Tin plate, terne plate, etc.	134.5
Ingots, blooms, billets, sheet bars, and skelp	104.2
Black steel sheets	100.5
Galvanized sheets	75.0
Iron and steel bars	53.5
Steel rails	51.7
All other products	440.3
Total	3,063.7

* Unit: thousand long tons (of 2,240 lb.). Source. "Statistical Abstract of the United States, 1936," Washington, U. S. Department of Commerce, 1936, p. 706. (Classes with under 50 thousand tons have been combined.)

TABLE 3

STEEL INGOT PRODUCTION IN THE UNITED STATES, MONTHLY, 1937*

January	4,725	July	4,556
February	4,414	August	4,876
March	5,216	September	4,298
April	5,070	October	3,393
May	5,150	November	2,154
June	4,184	December	1,472

* Unit: thousand long tons. Source: *Survey of Current Business*, February, 1938, p. 48.

Series classified according to some qualitative characteristic—such as the geographical location in Table 1 or the type of product in Table 2—are called *categorical series.* A very wide variety of such series exists: for example, a series in which daily wage rates of workers are classified according to their occupations, a series in which prices are classified according to type of commodity, a series in which the values of new buildings are classified according to the purposes for which the buildings were erected, a series in which amounts of tax collected are classified according to the commodities or activities upon which they are levied.

For Tables 3 to 6, on the other hand, the rule of classification is in terms of a measurable magnitude. Tables 3 and 4, in which the rule of classification is expressed in dates, present *time series;* in Table 3 classification is according to intervals of time and in Table 4 it is according to isolated dates. Tables 5 and 6 give *frequency series,* series in which the various observed objects have been classified according to their size, and in which the number of cases—frequency—for each size class is tabulated. In Table 5 each size class is stated as an *interval:* the size of any case counted as in that interval may, so far as we know, fall anywhere within

TABLE 4

TREASURY DEPOSITS WITH FEDERAL RESERVE BANKS AT END OF EACH MONTH IN 1937*

January 31	195	July 31	233
February 27	194	August 31	139
March 31	311	September 30	141
April 30	88	October 30	114
May 31	73	November 30	121
June 30	93	December 31	142

* Unit: million dollars. Source: *Federal Reserve Bulletin*, March, 1938, p. 198.

TABLE 5

NUMBER OF UNITED STATES CLASS I RAILROADS EMPLOYING STATED DEPRECIATION PERCENTAGE RATES UPON FREIGHT-TRAIN CARS IN 1936*

Rate (%)	Number of roads	Rate (%)	Number of roads
2.00–2.49	2	5.00–5.49	2
2.50–2.99	18	5.50–5.99	2
3.00–3.49	41	6.00–6.49	1
3.50–3.99	42	6.50–6.99	4
4.00–4.49	14	7.00–7.99	
4.50–4.99	5	8.00–8.99	1

* Source: "Statistics of Railways in the United States, 1936," Washington, Interstate Commerce Commission, 1937, p. S-80.

TABLE 6

NUMBER OF CONSOLIDATED CORPORATION TAX RETURNS IN UNITED STATES IN 1933 HAVING STATED NUMBERS OF SUBSIDIARIES*

Number of subsidiaries	Number of returns	Number of subsidiaries	Number of returns
1	3,638	11	53
2	1,199	12	51
3	608	13	36
4	360	14	39
5	260	15	28
6	154	16	26
7	128	17	26
8	93	18	20
9	80	19	12
10	50	20 or more	218

* Source: "Statistics of Income, 1933," Washington, U. S. Treasury, 1935, p. 35. (Twenty-two returns, with number of subsidiaries not reported, are excluded. Classes above 19 have been combined.)

the interval. Such a frequency series is called a *grouped* frequency series; it is also sometimes called *continuous*, although strictly this term applies to the variable and not the series. In Table 6, however, each size class is stated as an exact magnitude—an exact

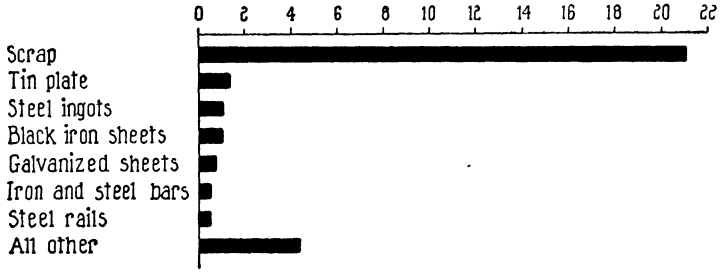


CHART 1.—Exports of heavy iron and steel products from the United States in 1935.
(Unit: hundred thousand long tons. Data in Table 2, p. 13.)

number of subsidiaries, in this case—and we know that each consolidated return counted in a particular class has the exact number of subsidiaries stated for that class. Such a frequency series is called *discrete*.

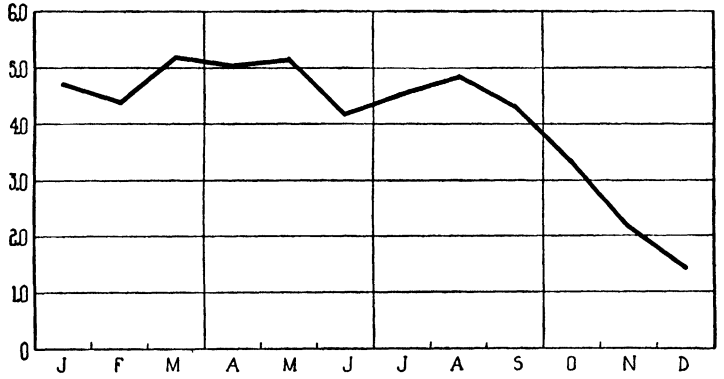


CHART 2.—Monthly production of steel ingots in the United States in 1937.
(Unit: million long tons. Data in Table 3, p. 13.)

Graphic presentation.—Although the tabular arrangement is the fundamental form for presenting a statistical series, a graphic representation—in a chart or diagram—is often of great aid in the study and reporting of statistical facts. Moreover, sometimes statistical data must be taken, in their sources, from graphic rather than tabular records. The technique of constructing and inter-

preting statistical charts requires careful study, and will be discussed below (Chaps. VII to IX). For the present, Charts 1 to 4 are shown merely as simple illustrations of the graphic method.

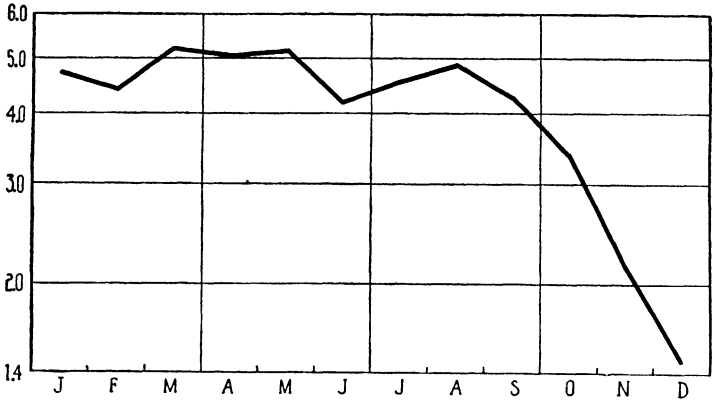


CHART 3.—Monthly production of steel ingots in the United States in 1937.
(Unit: million long tons. Data in Table 3, p. 13. Logarithmic scale.)

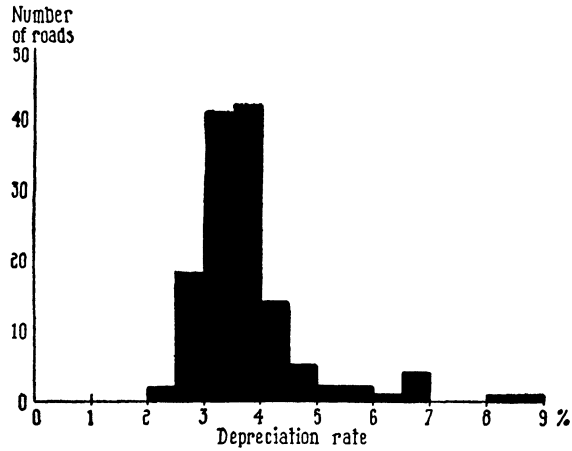


CHART 4.—Number of Class I railroads in United States in 1936, classified by depreciation rate upon freight-train cars.
(Data in Table 5, p. 14.)

Chart 1 is a bar diagram, representing the categorical series of Table 2. Charts 2 and 3 represent the time series of Table 3 in two different forms: in Chart 2 an *arithmetic* vertical scale is used, and in Chart 3 a *logarithmic* vertical scale. As will be explained below

(Chap. VIII), the second of these plans facilitates comparison of percentage changes in the observed variable (production). Chart 4 is a *block diagram* (also called *histogram*), which presents graphically the frequency series of Table 5.

. . . . . .

The tables and charts shown above are mere illustrations—and very simple illustrations—of the tabular and graphic methods of presenting statistical data. Later chapters will discuss at length the significant features of tables and charts, their construction and interpretation. Only after the student has acquired extensive understanding of these elementary aids to statistical work will he be ready to proceed with the more intricate arithmetical operations which form the central body of so-called analysis in statistics.

CHAPTER II

VARIABLES AND HOMOGENEITY

ECONOMIC VARIABLES

Statistical science, as applied in economics or any other field of inquiry, is almost without exception concerned with comparisons. As already noted, in certain rare and exceptionally simple instances the only numerical fact required as evidence may be a single isolated statistical item. But such instances are so unusual that they may be ignored in making the general assertion that statistical analysis treats of bodies of statistical data and their comparison and summarization. Statistical series, rather than isolated statistical items, form the factual basis of statistical analysis. Comparison of the items of a series, or of several series, is a central objective of statistical study.

In even the simplest study in economic statistics, therefore, we are concerned with an economic magnitude capable of taking on several different sizes—the various sizes being reflected in the various items of some appropriate statistical series. Such a magnitude, capable of taking on different sizes, is called a *variable;* and each of the several particular sizes is called a *variate.*

Variation in time and otherwise.—The term *variable* unfortunately tends to convey to our minds the notion of change over time—we unconsciously infer that the different sizes which the variable is capable of taking on occur at different times. It is true, of course, that some variation does occur over time. In such a case, the several variates are recorded in a statistical *time series.* But we must get rid of the notion that the words *variable* and *variation* necessarily imply change from time to time. They also refer to mere *differences* at a given time—differences between place and place, differences among persons, differences among objects, or differences among any other entities properly defined or described. Variation of this sort, from which the time element is absent, is represented in its simplest form by a *categorical series.*

The *frequency series* is a special case of representation of statistical variation from which the time element is ordinarily absent. The peculiar feature of the frequency series is that the variate appears in the rule of classification—generally the stubs of

the table—rather than in the numerical item itself. In the time series or the categorical series, on the contrary, the variate is the numerical item itself. In the frequency series, the numerical item states merely the number of times—frequency—of occurrence of a variate of particular size (or within a specified size interval). Thus, in Table 2, the first item—2,104—is a variate, a particular size of the variable; and likewise in Table 4, the first item—195—is a variate; but in Table 5, the first item—2—merely states that there are two variates between 2.00 and 2.49. To repeat: in a *time* or *categorical series*, the *variable is represented by the numerical items* themselves; in a *frequency series*, the *variable is represented by the rule of classification*—the succession of size classes.

As already indicated, the time element is ordinarily absent from a frequency series. The variable—represented by the size classes—customarily refers to differences among persons, objects, or other observed cases, at one particular time. Occasionally, however, frequency series are studied for which the variable—represented by the size classes—refers to changes over time.

For example, if we had a record of monthly steel ingot output, analogous to that of Table 3 but extending over the 144 months from 1926 to 1937 inclusive, we could classify those 144 variates according to size and thus obtain a frequency series. Table 7 shows the *time series* for the 144 months, with each variate recorded as a separate statistical item; and Table 8 shows the corresponding frequency series, with the number of instances (months) in which variates occur within particular size intervals recorded in the corresponding statistical item.

And here appears an important fact about a frequency series: *the identity of a particular variate is lost in the frequency series*. This is true whether the frequency series classifies a variable representing changes in time (as in Table 8) or mere differences without any time element (as in Tables 5 and 6). *In a time series or categorical series*, on the other hand, *the identity of each variate is preserved*—it is recorded in a separate statistical item. Thus in Table 7 the variate for October, 1930, is definitely stated as 2,693 thousand tons; whereas in Table 8 this variate is buried among the twenty cases falling in the interval between 2,500 and 3,000 thousand tons. Of course, for those classes in a frequency series—there is no such class in Table 8—for which the frequency is 1, the particular variate remains alone and is not mingled with others; but even here we should not know from the frequency table the date of that variate—its identity is lost.

TABLE 7

STEEL INGOT PRODUCTION IN THE UNITED STATES, MONTHLY TOTALS*

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1926	4,132	3,785	4,469	4,106	3,928	3,734	3,635	3,987	3,913	4,074	3,706	3,467
1927	3,823	3,845	4,575	4,163	4,083	3,526	3,232	3,529	3,298	3,345	3,155	3,203
1928	4,028	4,081	4,549	4,345	4,246	3,778	3,841	4,217	4,186	4,693	4,306	4,055
1929	4,500	4,329	5,068	4,950	5,286	4,903	4,851	4,939	4,528	4,534	3,521	2,903
1930	3,778	4,035	4,254	4,109	3,983	3,419	2,922	3,061	2,840	2,693	2,212	2,080
1931	2,512	2,547	3,054	2,767	2,552	2,128	1,888	1,717	1,545	1,591	1,592	1,301
1932	1,485	1,481	1,433	1,260	1,125	913	807	847	992	1,087	1,032	861
1933	1,017	1,073	898	1,345	1,976	2,564	3,168	2,864	2,283	2,085	1,521	1,799
1934	1,997	2,212	2,798	2,936	3,399	3,059	1,489	1,381	1,269	1,482	1,611	1,964
1935	2,870	2,774	2,865	2,641	2,634	2,259	2,268	2,916	2,825	3,143	3,150	3,073
1936	3,046	2,964	3,343	3,942	4,046	3,985	3,923	4,195	4,161	4,545	4,337	4,432
1937	4,725	4,414	5,216	5,070	5,150	4,184	4,556	4,876	4,298	3,393	2,154	1,472

* Unit: thousand gross tons. Source: for 1926–1935, Standard Statistics Company's *Basic Statistics*, Sec. G, June 5, 1936, p. G9; for 1936, *Survey of Current Business*, February, 1937, p. 48; for 1937, *ibid.*, February, 1938, p. 48.

TABLE 8

FREQUENCY DISTRIBUTION OF THE 144 MONTHS, 1926–1937, ACCORDING TO SIZE OF STEEL INGOT OUTPUT*

Lower limit of size class[a]	Number of months
500	6
1000	16
1500	11
2000	9
2500	20
3000	18
3500	19
4000	26
4500	14
5000	5
Total	144

* Compiled from Table 7.
[a] Unit: thousand tons.

The frequency series, then, in placing its entire emphasis upon size, cuts loose from the descriptive classification in terms of time or category. As already stated, most frequency series, unlike that of Table 8, are based upon variables which reflect not changes over time but rather differences at some particular time. The theory of frequency series, as we shall see, is developed in terms of this commoner type of frequency series, that in which the variates

reflect mere differences at a given time; and certain special adaptations of the theory are needed for the less common type of frequency series, that in which the variates reflect changes over time.

Complexity of variables.—The variables which are encountered in economic statistics are, in the great majority of practical problems, not simple and elementary but more or less complicated composites. By industrious search we can, of course, find some instances of simple or elementary variables; but the very difficulty of finding such instances emphasizes their rarity, and most instances which at first glance appear to be of this sort are found upon examination to have in fact a composite nature.

For example, a categorical list showing the number of bushels of wheat, of a specified grade, sold on specified terms of sale in each separate transaction at the Chicago Board of Trade during a specified trading day appears to be a series having an elementary rather than a composite nature. Likewise, a time series recording the number of tons of anthracite, of a specified size and quality, raised at a single shaft of a single mine on each working day for some specified period of time appears to be elementary. Also, a frequency series showing the number of cattle, of specified quality, according to classified weights, delivered to a railroad shipping point by a specified cattle raiser in a single specified shipment reflects an elementary variable.

But these cases, which were discovered only by considerable effort, pertain to situations which are so narrowly defined and so specialized that they are of little use for studying any significant economic problems. They do in fact assist in numerically describing particular elements in the economic system. Only a moderate relaxation of the narrow specialization of the above cases is needed to yield variables which are not elementary. Thus, the number of bushels of wheat, of the specified grade, sold on the specified terms of sale at the Chicago Board of Trade during the stated day by each of several traders—assuming that he does or may participate in several transactions—is not an elementary variable. For each trader, the total sales is a composite of his sales in separate transactions, which may be in various amounts, at various prices, and to various purchasers. Likewise, the daily tonnage output of anthracite by a company which operates several mines, or even several shafts at a single mine, is a composite variable. And the frequency series of weights of cattle shipped by a group of producers does not have the elementary character of the series for the single producer.

Economic science is concerned mainly with the study of situations and conditions which are not narrowly particular and specialized. Economics is chiefly concerned with simple and complex commodities in the mass, with commodities which are complex because of fabrication and services which are compounded of numerous functions, with processes which often involve numerous and intricate operations, with factors of production and agencies of consumption which are made up of more or less distinguishable parts, with complicated institutions and systems which carry on the work of production and distribution, and with the ramifying interrelations of all these elements. In the nature of economic problems, therefore, lies the reason for the inevitably composite character of most economic variables. Although the student may, by diligent search, find a moderate number of simple or elementary economic variables, and although careful study of the more common composite variables often requires an attempted expression in terms of elementary variables, the great bulk of the variables actually encountered and studied in economic problems must be regarded as composite. The following paragraphs describe briefly some types of composite economic variables.

Aggregates.—Many economic variables are mere aggregates—each is the sum or total of elementary economic variables of the same nature. Thus, steel ingot production (Table 7) is an aggregate of the production by various steel companies in their various plants and in the various furnaces of the several plants. Moreover, the monthly figure may be regarded as the total of daily figures, or even—if we wished to push the time recording that far—of hourly figures. Perhaps we should agree that the elementary variables, many of which when combined yield the composite variable reported in Table 7, are the outputs each day of each particular steel furnace. Several of the other variables illustrated in the foregoing tables are aggregates—in Table 1 each increase in loans is an aggregate of constituents for different banks in the group of banks specified by the stub and of different specific loans for each bank; each export item of Table 2 is an aggregate of numerous specific shipments; and each item of Table 4 is an aggregate of the specific Treasury deposits at the several reserve banks.

Instances of variables which are aggregates are very numerous and appear in cases in which the economic concept involved seems at first glance fairly simple. Thus, the cost of production—whether of a simple commodity like a bushel of corn or an elaborate

commodity like a locomotive—is an aggregate of more or less numerous specific costs (subject to the adequacy of accounting techniques) for materials of various sorts, for the use of land, for the use of equipment of various sorts, and for various sorts of labor and superintendence. Likewise, the value of a particular piece of real estate is a sum of various constituents—value of the land in one or more parcels, value of the building stated perhaps in terms of the specific values of various parts or features of the building. The weekly pay roll of a cotton-spinning factory is an aggregate of the wages paid to workers in each particular occupation and, more fundamentally, of the wages paid to each individual worker. The consumption expenditures of a specified family in a specified week are an aggregate of amounts spent for numerous items of food, clothing, fuel, shelter, amusement, and other types of consumption. The freight tonnage moved by a particular railroad in a particular month is an aggregate of various specific commodities, of different sizes and weights, hauled different distances, paying different rates, and involving different costs of transportation. The list can be extended indefinitely with ease, but enough cases have been cited to show how common the aggregate is as a type of composite economic variable.

Rates and ratios.—Another important type of economic variable appears as a rate, or ratio. This is obviously a composite in that the given variable is the quotient of two other variables, the numerator and the denominator. For example, the *operating ratio* of a railroad is the quotient of its operating expenses, in a given interval such as a specified month, divided by the operating revenues. Here variation in the ratio—for example, from month to month—is manifestly the composite, or resultant, of variations in both numerator (expenses) and denominator (revenues). Likewise, the *rate of turnover* of inventory in a retail store in a specified year is the quotient of the sales during the year by the value of the stock of goods (expressed as an average, because the stock varies during the year) held for sale during the year. Here variation in the rate—for example, from store to store—is the composite, or resultant, of variations in both numerator (sales) and denominator (value of stocks).

The same is true of a long list of economic variables which, though they may not include in their titles the word *ratio* or *rate*, are in fact mere quotients. Thus, important examples include: (1) *reserve ratio* of a bank, (2) *traffic density* of a railroad, (3) *net worth ratio* of a corporation, (4) *velocity of circulation* of money,

(5) *rate of turnover* of labor, (6) *yield* of a bond, (7) *markup* of retail merchandise, (8) *yield per acre* of a crop, (9) *factor of safety* of a bond, (10) *current ratio* of a business firm, (11) *earnings ratio* of a corporation, (12) *productivity* of labor, (13) *unit cost* of output of a commodity, (14) *load factor* of a power plant, and (15) *turnover rate* of receivables.[1]

Certain economic variables are in fact rates, or ratios, though we do not customarily think of them as quotients of a numerator by a denominator. Thus, *price* is strictly a ratio: the quotient of the money paid divided by the amount of the commodity received. Likewise, the *rate of interest* is the quotient of the amount of the interest divided by the amount of the principal. A *wage rate* is the ratio of the amount of wages to the time worked. An *exchange rate* is the ratio of the amount of one currency to that of another for which it is exchanged. A *depreciation rate* is the quotient of the loss of value—usually estimated in physical terms, with respect to "useful life"—of a machine (or other capital instrument) during a year divided by its value at the beginning of the year, or more generally at the beginning of its use. A *freight rate* is the ratio of the charge for transportation between two points to the volume, or weight, hauled. A *tariff rate*, of the ad valorem sort, is the ratio of the duty paid to the value of the imported article. A *tax rate* is the amount of tax levied, divided by the tax base—usually the value, but sometimes the quantity, of the thing taxed. A *commission rate* is the ratio of the charge for selling to the value sold. This is part of a long list of economic variables which are in fact ratios but which are customarily treated as if they were elementary—rather than composite—variables. Such treatment is the consequence partly of legal arrangements which express business and economic relations in terms of these ratio concepts, and partly of mere custom in the conduct of economic affairs.

For these rates, as for the more obvious type of rates (like types 1 to 15 above), the variation of the ratio is a composite, or resultant, of possible variations in both numerator and denominator. In this second class—which includes rates such as price, wage rate, tax rate—we find it less easy to think of numerator and denominator varying independently of each other, than for the first class—which includes such rates as reserve ratio or velocity of circulation

[1] The definitions of some of these concepts are obvious, but the concepts are defined and explained in textbooks dealing with banking, railway operations, corporation finance, etc.

or earnings ratio. In the second class, we are in the habit of thinking of the rate as *given*, and the "numerator" thus appears merely as the "denominator" multiplied by the rate. But a moment's thought will show that, if the rate is not *fixed*, both numerator and denominator can vary separately.

In studying ratios of all sorts, in fact, we have to consider that both numerator and denominator can vary, though recognizing that the degree of independence of the variations of the two elements differs from case to case. In a case where the rate is legally fixed, as for a tax rate of a particular kind at a particular time, dependence of numerator upon denominator is practically complete. Near the other extreme is the case of the *current ratio*—the ratio of current assets of a firm to current liabilities—for which a variation in the numerator is to a very important degree, though not wholly, independent of a variation in the denominator.

A special class of rates arises in connection with all sorts of per capita figures. In all such cases, some aggregate (numerator) is divided by the corresponding number of people (denominator). Examples are: per capita income, per capita expenditures on amusements, per capita output of soap.

The per capita rates suggest an important general point. A per capita figure is obviously an average: per capita income is the average income per person. In general, for any rate or ratio, the average idea is implied: thus, *acre yield* of corn on a farm is the average number of bushels per acre. The theoretical importance of this point is very great. As we shall see (Chap. X), the significance of a statistical average depends largely upon its typicalness; and the same is accordingly true also of a rate or ratio. The full implications of this assertion cannot, however, be brought out until we have explored the theoretical foundations of averages.

Variables as products.—Just as an economic variable may be a ratio of a variable numerator to a variable denominator, a type of composite variable exists which implies multiplication of two or more elementary variables. For example, the variable representing performance of railroad transportation is the number of *ton-miles*—obtained in multiplying the weight hauled by the distance covered. A variable measuring performance of equipment in a factory is the number of *machine-hours*—obtained in multiplying the number of machines in operation by the time operated. A similar measure for labor performance is stated in *man-hours*.

Similarly, other composite variables may result from combining pairs or groups of variables by other more or less elaborate arith-

metical formulas. The arithmetical operation of dividing merely happens to be the most common and that of multiplying the next most common. And, of course, several arithmetical operations may be used in making one composite variable. For example, *traffic density* (defined above, page 9) is a ratio in which the numerator also involves a multiplication. Also, *rate of turnover* of inventory (defined above, page 23) is a ratio in which the denominator may be the result of the statistical process of averaging.

Derived statistical variables.—Another important type of composite economic variable takes the form of a calculated statistical number. Examples are: all sorts of statistical averages, measures of variability, index numbers, correlation coefficients. These concepts are too intricate to admit of brief explanation at this point, and will be discussed in later chapters. They are mentioned here, however, to emphasize the important fact that even an elaborately calculated statistical summary figure, derived from economic data, becomes at once an economic variable, capable of changes from time to time or of differences between places or cases. And always in the study of such an economic variable, as for any composite variable, we must bear in mind that its variation is the resultant of variations in its constituents.

HOMOGENEITY

We have seen above that many economic variables are composites, and an obvious implication of this complexity of a particular variable is that the study of such a variable—of statistics pertaining to that variable—must be in the light of the capacity of its constituents to vary, and of such constituent variations to be more or less independent of each other.

The statistician seeks not only to *describe* a variable by the use of statistics but also—though often less successfully—to *explain* the variation. Ultimately, such explanation aims at discovering causes or possible causes of variation, but this phase of explanation so clearly involves theoretical considerations that the work of the statistician generally stops short of assigning causes. What he generally does, rather, is to explain—perhaps only partially and tentatively—the variations of one variable in terms of those of other variables. Clearly, for those cases in which a given variable is a composite of other variables this kind of explanation first aims at relating statistics of the changes or differences in the given variable to those of its constituent variables. This method is one

of *analysis*, in its literal sense: the breaking up of a whole into its parts in the expectation that knowledge of the parts will afford knowledge of the whole. The complexity of the typical economic variable thus suggests a useful approach for studying statistics reflecting economic variation.

But this complexity has another practical implication: it increases the danger that a statistical series, reflecting an economic variable, will not be homogeneous. The term *homogeneity*, as pertaining to an economic series, does not readily admit of definition; and an understanding of the term will perhaps best be secured by studying the circumstances or conditions which impair or destroy homogeneity. The difficulty of definition is enhanced, as will be observed in the following paragraphs, by the fact that the same statistical series may be tolerably homogeneous for one purpose and very far from homogeneous for another. In particular, the homogeneity here under examination has not the specialized connotations of that term as used in the theory of sampling.

A tentative definition can be stated in this form: A statistical series is strictly homogeneous, for a given purpose of study, if its various items accurately record the variations—over the interval of time or range of cases studied—of a single variable having a rigid and unchanging definition. This suggests that impairment of homogeneity may result from the facts that: (1) the statistics are not accurate, (2) they do not pertain to a single variable, (3) the variable does not have a rigid and unchanging definition. As indicated above, the homogeneity concept will be clarified by studying ways in which homogeneity may be impaired. In the cases cited below, frequent mention will be made of the extent to which the purpose for which statistics are used affects the question of homogeneity.

Changes in definition over time.—The definition of a variable, represented by a statistical time series, sometimes changes with passing time. Such changes are not always—perhaps not often—evident in the statistical record as presented in a table; and an initial task of the statistician is to test by every available means the homogeneity of his time series, from this point of view of the definition of the variable.

As a first example, consider the series in Table 9. It is a mistake to suppose, merely because the statistical item is described by a fixed phrase in the title of the table that the variable had an unchanging definition over these fourteen years. By consulting the source we find a long list of notes (pages 112–119) concerning

comparability of the figures from individual income tax returns. Some of these notes pertain to the definition of net income; among the most important of these are changes in the law and regulations as to treatment of capital gains and capital losses and as to the carrying forward of a net loss incurred in one year to later years. By examining these notes, the statistician at once learns that the variable reflected in Table 9 is not rigidly defined and that the series is therefore not homogeneous.

TABLE 9

AGGREGATE NET INCOME REPORTED ON INDIVIDUAL INCOME TAX RETURNS IN THE UNITED STATES*

1921	19,577	1928	25,226
1922	21,336	1929	24,801
1923	24,777	1930	18,119
1924	25,656	1931	13,605
1925	21,895	1932	11,656
1926	21,959	1933	11,009
1927	22,545	1934	12,797

* Unit: million dollars. Source: "Statistics of Income, 1934, Part 1," Washington, U. S. Treasury, 1936, p. 22.

Similarly, a series recording the number of individual income tax returns filed would be discovered to be lacking in homogeneity. Reference to the same source cited for Table 9 would show that various changes in the law, chiefly with reference to the minimum income for which a return must be filed, destroy all semblance of rigidity in definition of the variable "number of returns."

As another example, consider the annual average price of lead at New York in cents per pound during the period 1930–1937. Recalling that, until early in 1933, the currency of the United States had for many years been redeemable in gold at a fixed ratio, that during part of 1933 and early 1934 the relation between the currency dollar and gold was subjected to official manipulation, and that the Gold Standard Act of early 1934 and subsequent official proclamations "fixed" a new relation but not redeemability between currency and gold, we observe that the series for price of lead cannot be regarded as homogeneous. "Price" of lead—or of any other commodity having a price stated in dollars—is not rigidly defined during that interval. To be sure, if we are definitely interested in price in terms of United States *currency*, such a series may be treated as homogeneous; but as soon as we even by implication think of price in terms of gold or of foreign currencies, homogeneity disappears.

Table 10 shows the average number of employees of the United Steel Corporation and subsidiaries in each year since the war. This series is not, despite its appearance, homogeneous. Recalling that, shortly after the war, agitation for replacing the standard 12-hour shift then common in many branches of the steel industry by an 8-hour shift was successful, we examine the source cited in the table for evidence. The 1923 issue of the source

TABLE 10
AVERAGE NUMBER OF EMPLOYEES OF THE UNITED STATES STEEL CORPORATION AND SUBSIDIARIES, ANNUALLY*

1919	252	1928	222
1920	267	1929	225
1921	192	1930	211
1922	215	1931	204
1923	261	1932	158
1924	247	1933	173
1925	250	1934	190
1926	253	1935	195
1927	232	1936	222

* Unit: thousand employees. Source: *Annual Reports* of the U. S. Steel Corporation, 18th (1919) to 35th (1936)

shows (page 29) that the change went into partial operation in August, 1923, and into full operation in February, 1924; part of the effect of the change was felt in 1923 and 1924 and the full effect in 1925. The fact of this change does not render the series non-homogeneous for the purpose of recording changes in the size of the labor force of the corporation; but, as a measure of the amount of labor going into production or as an indicator of activity of the corporation, and for numerous other collateral purposes, the series is not homogeneous. Again, we find in the 1930 issue of the source (page 10) reference to the "staggering" of employment during depression, and subsequent issues give further textual and tabular evidence enabling us to study part-time employment by the corporation. This change in policy in 1930, and the developments of the new policy thereafter, must also be noted as an impairment of homogeneity. The acquisition of important new subsidiaries by the corporation during the period covered by the table must also be examined, along with other possible causes of nonhomogeneity not here mentioned, before the extent of the defects in the homogeneity of the series can be known.

As an example of the effects of acquisition of new property upon homogeneity, consider Table 11. The series of freight

TABLE 11
FREIGHT REVENUES OF THE NEW YORK CENTRAL RAILROAD COMPANY, ANNUALLY*

1927	234	1932	193
1928	235	1933	194
1929	242	1934	204
1930	307	1935	218
1931	246	1936	258

* Unit: million dollars. Source: "Statistics of Railways in the United States," Washington, Interstate Commerce Commission, issues for 1927 to 1936.

revenues is not homogeneous because on February 1, 1930 (see source, 1930 issue, page 24) the New York Central Railroad Company leased certain lines it had controlled by stock ownership, including the Michigan Central, Big Four, and certain smaller lines, and the operating revenues of the lessee thereafter included figures for the leased lines. Detailed examination of back issues of the source would enable us, by combining figures, largely—but not wholly—to restore homogeneity to the series.

Suppose a series of annual averages, since 1919, of the price of steel bars, quoted for delivery at Chicago is given. Recalling the long controversy concerning "Pittsburgh plus" in the steel industry, we examine collateral records and find that: for many years prior to 1922, various steel products, wherever manufactured, had generally but not invariably been quoted in each market on the basis of the Pittsburgh price plus freight charges from Pittsburgh to that market; that in January, 1922, this practice was abandoned for some products, including bars; and that a further partial abandonment occurred in July, 1924. Later, under the N.R.A., a modified form of the basing-point system of quoting prices was put into effect.[1] These changes in the manner of quoting steel prices affect the homogeneity of the series: it remains homogeneous as a mere record of the cost of steel bars to Chicago consumers, but it is not homogeneous for the purpose of analyzing price conditions in the steel industry.

Changes in definition, in categorical series.—Several illustrations, given herewith, will indicate the danger that a variable reflected by a categorical series, though it may appear superficially to be rigidly defined, can have serious defects in homogeneity. The series in Table 12, for example, is defective in homogeneity because the word *cotton* does not imply precisely the same commodity,

[1] ARTHUR R. BURNS, "The Decline of Competition," McGraw-Hill Book Company, Inc., New York, 1936, pp. 299–317.

from the point of view of quality and therefore of technical and commercial usefulness, in the various countries. The chief element in the quality of cotton is the length of staple; and staple length varies somewhat for cotton produced within each country, but differences between the countries as to average length are important. On the average, Egyptian staple is long, generally ranging from $1\frac{3}{16}$ to $1\frac{3}{4}$ inches; Indian and Chinese staples are short, mainly of $\frac{7}{8}$ inch or shorter; whereas the bulk of American cotton has staple between $\frac{7}{8}$ and 1 inch. Direct comparisons of the items in Table 12, without reference to these facts, are misleading.

TABLE 12

PRODUCTION OF COMMERCIAL COTTON IN LEADING COUNTRIES, IN 1936–1937 CROP YEAR*

United States	12,375	Egypt	1,863
India	5,661	Brazil	1,708
China	3,256	Other countries	2,587
Russia	3,250	Total	30,700

* Units: thousand running bales, for the United States; thousand equivalent bales of 478 lb. net weight, for foreign countries. Source: "Cotton Year Book of the New York Cotton Exchange, 1937," New York, p. 11.

A series giving the assessed valuation of real estate in each city of the United States is not homogeneous for the purpose of studying differences in physical wealth, although it is tolerably homogeneous for comparing differences in the "tax base"—the value against which a tax is levied—among cities. The essential point is that differences in law or in administrative practice result in differences, among cities, in the degree to which assessed valuation reflects "true" value. No satisfactory data exist, telling the percentage relation of assessed value to true value, in various cities, largely because the very definition and determination of "true" value are difficult or impossible. But experience teaches that such percentages, if they did exist, would vary widely from city to city; and we are warned therefore against treating the assessment data as homogeneous.

A series showing the principal property account—value of lands, buildings, and equipment—of several companies, as taken from their balance sheets, may not be homogeneous. Examination of the sources (balance sheets and related documents) may disclose that some of the companies report their property values at original cost whereas other companies give net figures after a deduction for depreciation.

Again, a series giving the physical property values, as stated in balance sheets, for several retail establishments may be defective in homogeneity for the purpose of studying differences in the amount of real estate and equipment used in the several businesses. Some firms may own all or practically all of the property used, whereas others rent a large share of the necessary plant and equipment. In the first case, the property would appear in the balance sheet; in the second, it would not. Careful study of the sources of data, and collateral evidence, will generally—though not necessarily—bring out these facts, and the user of such a statistical series must not fail to make this study.

Similarly, a series of balance-sheet figures on inventory, for a list of companies, may not be homogeneous, because some companies value their inventories at cost while others report cost or market value, whichever is lower. Particularly in a period when market prices are declining rapidly, the effect of disparity between cost and market values upon balance-sheet valuations may be very important. Here also, unless the facts are specifically recorded in the table giving the series, the user of the data has no choice but to discover the facts before he interprets or analyzes the statistics.

The series giving volume of passenger traffic in Table 13 is not homogeneous for many purposes. Examination of the source—as well as general information—shows that the typical passenger is hauled much farther on some lines, like the Burlington or the Atchison, than on others, like the New Haven or the Lackawanna. Moreover, the passenger business of some lines, like the Long Island or the Jersey Central, consists very largely of commuters; whereas commutation traffic is negligible on other lines, such as the Southern or the Union Pacific. These facts impair the homogeneity of the series for many studies, including those of rates, costs, need for equipment, maintenance of equipment and plant, operating employment and pay roll, stability of revenues, and the like.

A series giving, for a particular time, the number of manufacturing companies in Belgium, England, France, Germany, Italy, and the United States would not be homogeneous. Differences between countries—as to laws, customs, and public policy—control in an important degree the extent to which an industry is organized in the company form, and the very meaning of the term *company*. Moreover, for any one country, a mere count of the number of companies ignores the great range in size among particular companies. Though each item of the supposed series wears

TABLE 13
REVENUE PASSENGER-MILES IN 1935 FOR SELECTED CLASS I RAILWAYS OF THE UNITED STATES*

Boston and Maine	318
New York, New Haven, and Hartford	1036
Delaware, Lackawanna and Western	424
New York Central	2246
Baltimore and Ohio	460
Long Island	1277
Pennsylvania	2217
Central of New Jersey	333
Atlantic Coast Line	301
Louisville and Nashville	349
Illinois Central	482
Southern	484
Chicago and Northwestern	554
Chicago, Milwaukee, St Paul, and Pacific	350
Atchison, Topeka and Santa Fe	748
Chicago, Burlington and Quincy	425
Chicago, Rock Island, and Pacific	360
Southern Pacific	1138
Union Pacific	367

* Unit: million passengers one mile Source: "Statistics of Railways in the United States, 1935," Washington, Interstate Commerce Commission, 1937, Section A, pp. 16–125.

the same name as the others, it manifestly has a meaning peculiar to itself; and no careful statistician treats such a series as homogeneous.

A series purporting to record, for a given year, the national income of each of several countries is not homogeneous. Comparability of such data is impaired by numerous basic factors which affect the size, structure, and significance of the national income—factors such as the size, racial, age, and other characteristics of the population; by the degree of urbanization, the extent of organization and mechanization of industry, the forms of property ownership, and other economic and social characteristics of each country. But, waiving these factors which impair homogeneity of the national income data, we find those data lacking in homogeneity merely on the basis of definition. An intricate concept such as national income is defined only with difficulty; and the estimates for various countries, despite some effort by estimators to use equivalent definitions, rest upon widely various definitions. This fact alone detracts from the homogeneity of this series.

Indexes of the cost of living, for each of several localities in the United States, do not constitute a homogeneous series for most purposes for which they are likely to be used. Even if each index

has been computed in the same way, according to a definition and formula fixed for all localities, the comparability is merely superficial. The essential fact is that the concept *cost of living* means different things with reference to different localities—because of differences in climate, structure of population, and numerous physical and social factors.

Other defects in homogeneity.—The foregoing illustrations, of time series and categorical series which are not homogeneous, have emphasized lack of uniformity or fixity in definition of the variable. This is a very common cause of nonhomogeneity, and frequently the easiest cause to discover. The student will, as he acquires experience and skill, become acquainted with other and more subtle causes of nonhomogeneity and develop some understanding of where and how to look for them. As already noted, no rigid definition of the term *homogeneity* can be given; and no positive criterion can be set up by which, in a particular case, the homogeneity of a series can be tested once for all. The process of testing consists rather in a ramifying, but not therefore unsystematic, search for causes which *might* impair homogeneity.

Moreover, we have seen that homogeneity is a relative term: for some purposes a series may be tolerably homogeneous while for others it is not. The process of testing is therefore influenced by the use to be made of the statistics. The particular use will often suggest the possible causes of nonhomogeneity; and, in fact, scientific interpretation of a statistical series—by making comparisons or by other sorts of analysis—often uncovers causes which limit its homogeneity.

In the discussion of lack of rigidity of definition, we have made no mention of fixity of the unit in which the variable is measured. This seems to be an obvious requirement, and yet the sober fact is that some so-called statistical series do state the different items in different units. No excuse exists, of course, for failing to insist upon a uniform unit—or at least upon full knowledge as to any differences in the unit.

Likewise, the foregoing illustrations—except possibly in the property assessment (page 31) and national income (page 33) cases—have not stressed the *accuracy* of the statistics as a source of nonhomogeneity. Any evidence that particular items of a series are imperfect numerical observations of the variable, especially if the imperfection varies from item to item, must be weighed as bearing upon homogeneity.

One of the requirements stated above (page 27) for a homogeneous series is that the items pertain to a single variable. Actual cases in which all the various items of a series do not pertain to the same variable are fortunately infrequent where good tabulation techniques are used, and can generally be recognized at sight. A more elusive difficulty arises, however, in any case involving a composite variable; and this difficulty has to do with interpretation rather than with the intrinsic quality of the series. Frequently, in studying a composite variable, we are prone to regard its values as indicative not only of changes in the variable itself but also of changes in one or more of its constituent variables. Strictly, this does not have any bearing upon the homogeneity of the series itself; but, as the purpose for which a series is used governs somewhat its homogeneity, this point is mentioned in the present connection.

A wide range of cases in which series are not homogeneous, which the foregoing illustrations do not emphasize, arises because of lack of rigidity of definition of the *object* observed rather than of the variable itself. The individual net income case (page 28) is of this sort in one respect: changes in the requirement for filing returns (related to changes in the exemption) result in a different coverage in different years. The number of people liable to file returns changes from year to year—the *object* observed, the total body of people filing tax returns, is not rigidly defined.

Another example of this sort is a series of annual steel output by Germany from 1910 to 1936: changes in the national boundaries and political domain of Germany—particularly with reference to the Ruhr area—imply changes in the "object" under observation. In a sense, this is still a question of "definition of the variable," but cases in which the "coverage" is not rigidly fixed occur so commonly that some advantage flows from regarding such cases separately.

Homogeneity in frequency series.—As a frequency series merely organizes the several values of the variable—variates—in a special arrangement according to size, any of the causes which can impair the homogeneity of a time or categorical series can also limit the homogeneity of a frequency series. The fact that a frequency series conceals the identity of the variate, however, obstructs testing of homogeneity by the procedure indicated above. This does not mean that such procedure is futile and should not be attempted, but it does mean that causes of nonhomogeneity, which

are suspected for a particular frequency series, can sometimes not be directly examined. And the very form in which a frequency series appears removes somewhat our disposition to suspect such causes, gives us rather an insidious prejudice in favor of homogeneity.

Fortunately, on the other hand, certain characteristics of the form of distribution of frequencies according to size of variate afford presumptive evidence as to the homogeneity or lack thereof. We shall see that lack of homogeneity in a frequency series often arises from the inclusion, among the cases tabulated, of different groups of objects for which the variate in question has naturally different average sizes. But the whole question of homogeneity is peculiarly involved for frequency series; and, beyond emphasizing at this point that the student must use every known means for testing the homogeneity of such series, we postpone further comment until the methods of analyzing frequency distributions are discussed.

Homogeneity and comparability.—The great importance of homogeneity, as a characteristic of a statistical series, is that it is essential for comparability. Except in the rare instances in which isolated statistical items are used merely to give numerical reality to an economic concept, the use of statistics in economic analysis—and in other branches of science—involves and aims at comparisons among items. In making scientific comparisons a fundamental rule is that we must compare similar things.

Comparison among the items of a statistical series which is not homogeneous obviously defies this rule. Of course, the things compared are not exactly similar in all respects, else they would be exactly alike even as to size of the variable under observation, and there would then be nothing to compare. Unless the items differ in magnitude—unless they reflect a variable—there is nothing to be compared. What we really have in mind, in stating the fundamental rule, is that the cases or objects must be similar *except* as to the size of the variable under study. A series which is homogeneous meets this requirement: comparisons among the items of such a series really tell us something about the variable in question, instead of giving us a confused picture of the joint effect of several variables.

This manner of stating the case explains why homogeneity is so difficult to define and so difficult to determine in practice; why it depends upon the purpose for which the data are used, on the sort and significance of the comparisons to be made; and why it is so

seldom fully realized in economic data. Economic data, because of the composite character of most economic variables and because of the complex nature of most economic objects—such as families, commodities, companies, nations, industries, machines, plants, and even individual men and women—are inherently of the sort not amenable to description in terms of a single and truly elementary variable. Comparisons among such data are accordingly liable to reflect the effects not of a single and truly simple type of variation but of a multiplicity of variable causes.

In view of this circumstance, much of the method of statistics is directed to furnishing a means of analyzing *complex* variation, of breaking such variation into its parts, and of revealing if possible the separate elements in such variation. To be sure, elementary comparisons, which merely state which of the items of a series are larger than others and how much larger, have some use in economic analysis by the aid of statistics. But this rudimentary process of comparison is sufficient in relatively few cases, and chiefly only in those cases for which the statistical series are highly homogeneous. More elaborate analytical methods are, of course, of value even for analyzing such highly homogeneous series; and we shall find it helpful to begin and carry forward the study of such methods with reference to tolerably homogeneous series. But an important purpose of analytical methods is to provide some means, however imperfect, for understanding variation which is not homogeneous. All this is said without any intent to minimize the desirability, amounting almost to absolute necessity, of ensuring that every statistical series is made as nearly homogeneous as possible. The best raw material we can possibly obtain, for statistical analysis, will be none too good; and, before undertaking analysis, the student should exhaust every means of rendering his data homogeneous.

CHAPTER III

SOURCES AND THEIR USE

PRIMARY AND SECONDARY DATA

The great bulk of statistical analyses in economic investigations rest upon statistical material copied from published sources and therefore not originated for the immediate purpose of the investigation in hand. Such statistics are called *secondary data.* Because secondary data are the chief materials of statistical work in economics, problems connected with their use are the main subject of this chapter.

On the other hand, statistical material which the investigator originates for the purpose of the investigation in hand are called *primary data.* The process of assembling primary data is called *collection* of statistics. The student may be tempted to apply the term *collection* to the process of compiling statistics (secondary statistics) from various published sources, but he should note that the term is used strictly in the narrow sense defined above. Collection means the assembling, for the purpose of a particular investigation, of entirely new data, presumably not already available in published sources.

As noted above, most statistical analyses rest upon secondary data, and the statistical investigator in economics will therefore ordinarily make little use of primary data. He takes his numerical materials mainly from existing compilations, and frequently he makes no use of primary data. One reason why he relies so largely upon secondary data is that, particularly in recent times, the amount and diversity of published statistics are so great that he finds more or less ready to hand the main materials needed for studying numerous economic problems. A second reason is that because of the great complexity of many economic questions and the resultant necessity of securing data on a grand scale the mere burden of labor and expense stands in the way of extensive collection of primary data by the individual investigator.

The economic statistician needs, however, to have some familiarity with the problems of collecting primary data. He may, in a particular investigation, discover that secondary data do not

afford an adequate basis for his analysis and that he *must* secure some of his material by collecting primary data. To conduct this operation wisely, he needs to understand at least the principal difficulties likely to be encountered in such collection. He needs to know how to plan and conduct a survey directed to collecting primary data along lines which will effectively yield the desired information. Moreover, he may in the course of his professional work be attached to an organization or participate in an undertaking which has as one of its main objectives the collection of primary data. The purpose of an investigation, in the work of certain organizations, is precisely the collection and publication of primary data. Such is, in a broad sense, the case with the Bureau of the Census and various other governmental agencies, and, to a less degree, with certain agencies outside the government. The statistician who is to take part in such an enterprise needs sound understanding of the problems of collection and of the methods which have proved most effective in collection.

But there is a deeper reason why even the statistician who seldom or never uses any but secondary statistics should understand the collection process. Ultimately all secondary data which he uses rest upon a collection operation: somebody, sometime, must have brought those data into existence by a process of collection. Hence, true insight into the significance and limitations of secondary data requires knowledge in detail of the methods used in their collection by the person or agency which originally collected them as primary data. Without such knowledge, the user of secondary data remains in uncertainty as to many factors, some of which may be highly important for his purpose, affecting the quality and adequacy and other characteristics of the raw materials for his investigations. Because of the great importance of the collection process, even to the statistical investigator who does not himself collect primary data, the next chapter is devoted to a discussion of certain leading aspects of the collection problem.

PRIMARY AND SECONDARY SOURCES

In taking secondary data from a source, the initial question to be answered is whether the source originated the data—collected them as new—or merely quoted them from still another source. If the source which publishes the data collected them—if the statistics are primary data from the point of view of that source—that source is called a *primary source*, or sometimes an *original*

source. If the source merely quotes the data from some other publisher—if the statistics are secondary data from the point of view of the source—the source is called a *secondary source.* The first and general rule in compiling secondary data is to use the primary source. Assuming that we are not using primary data in an investigation, we seek secondary data in one or more sources. If we discover them in some particular source, we then inquire whether that source is primary for those data. If we discover that it is not the primary source, the rule requires us to go back of it until we finally locate the data in their primary source. This is the general rule; we shall note some exceptions and qualifications, but they are so slight that the student should almost invariably apply this general rule rigidly.

The rule is not essentially different, in its purport or its justification, from the same rule which applies to the use of all evidence, whether statistical or other. It is the rule of insisting upon first-hand evidence, of excluding so far as possible the second-hand and hearsay elements. It applies in every branch of scientific inquiry, and should require no special emphasis with respect to statistical inquiry. Perhaps the disposition to ignore or violate it is peculiarly present in statistics because of the attractive accessibility of data in numerous secondary sources. But, whatever the temptation to accept data from a secondary source, the student does well to insist obstinately upon primary sources. By so doing he will protect himself from many errors and imperfections in the raw material of his analyses.

Chief advantages of primary source.—The main reasons for insisting upon the rule are two. A primary source is more likely to describe carefully the process of collection than a secondary source. The user of a primary source therefore has the advantage of information concerning the collection process, emphasized above (page 39), and can appraise its adequacy and reliability for the purpose of his study. To be sure, a secondary source sometimes discusses and criticizes helpfully the collection process used by the primary source from which it quotes; and such a practice, whenever it is encountered, is of great aid to the user. But the practice is not common, and the user should not count upon finding it. Even when he does find it in a secondary source, he is all the more impelled to consult the primary source and form his own opinion as to the validity of the criticism.

The second main reason for the rule is that a secondary source cannot be relied upon to give an accurate and complete quotation

of statistical data from a primary source. Mere numerical errors in transcription may, and frequently do, impair accuracy in the secondary source. Here also a qualification is necessary: sometimes a secondary source discovers and corrects an error in the primary source, but these instances are not sufficiently frequent to warrant setting aside the rule on their account. Defects in completeness of quotation are far more serious than mere numerical errors. The secondary source often fails to reproduce all of the pertinent data from the primary source—perhaps merely in the interest of economy, but perhaps because the secondary source has a purpose in reproducing the data which leads it to omit or suppress some of them (see below, page 47). Moreover, the secondary source often fails to reproduce significant footnotes, or textual comments, by which the primary source had qualified the data or their definition or the units.

Sometimes a secondary source gives data in a different form from that of the primary source; for example, the secondary source may give the data in the form of percentages of some total or base figure, whereas the primary source gives the actual figures without any arithmetical conversion. Or, a secondary source may present merely a chart of the data, whereas they appear in tabular form in the primary source. In all such cases, danger that the process of conversion may have introduced errors, or otherwise damaged the accuracy or completeness of the data, is an argument for insisting upon the primary source.

Revised figures.—Systematic use of the primary source is more likely to reveal revisions in the data previously published. This practice of revising previous data most commonly arises, of course, in connection with time series, and particularly in cases in which successive periodical issues of a primary source publish the current data of a time series and republish, with revisions where necessary, earlier data. For example, Table 14 is an abstract of series from the *Federal Reserve Bulletin,* which is the primary source for these production indexes. The items marked *r* have been revised from earlier issues of the periodical; in the February, 1938, issue (page 152) they were reported as 89 and 100, respectively. Obviously, quoting data for these production indexes from some secondary source which, for one reason or another, had not "caught" the revisions would result in errors in the data used by the investigator.

Revisions of data in a primary source are not always so plainly marked as in the source quoted in Table 14, the *r* symbols in that case are quoted directly from the source. For example, the

Commercial and Financial Chronicle, which is a primary source for data on security flotations, gives in its issue of March 5, 1938 (page 1457), "municipal" new capital issues for February, 1937, as $33,504,423. By consulting the issue of March 6, 1937 (page 1501), we find the figure then currently given as $34,345,523. Obviously, unless there is a mere misprint or erroneous transcription in the more recent issue, there has been a revision of the February, 1937, figure during the year. Quotation of the February, 1937, figure from a secondary source which had not "caught"

TABLE 14

INDEXES OF INDUSTRIAL PRODUCTION IN THE UNITED STATES, MONTHLY FROM JANUARY, 1937*

Month	Total	Manufactures	Minerals
1937			
January	114	115	110
February	116	116	115
March	118	117	128
April	118	118	115
May	118	118	116
June	114	114	114
July	114	114	112
August	117	118	112
September	111	110	115
October	102	101r	113
November	88r	85	109
December	84	79	114
1938			
January	81p	76p	108p

* Unit: per cent of 1923–1925 average. Adjusted for seasonal variation. Source: *Federal Reserve Bulletin*, March, 1938, p. 222.
r revised; p preliminary.

this revision would supply the investigator with an erroneous item. Study of this primary source reveals frequent revisions of this sort: the monthly figures published currently are revised one year later, when they are published alongside the new data then current. A careful secondary source may include these revisions, but it is only by using the primary source that we can be sure of catching them.

Finally, cases occur in which the primary source makes revisions in previous data without any systematic republication of the old table with the newly revised items. In such instances we discover evidence that revisions have been made only by examining the textual material, or supplementary notes not necessarily attached to any table, in later periodical issues of the primary

source. For example, *Statistics of Income* for 1931[1] carries on pages 32 and 33 text notices of revisions of particular items in tables of issues of the same report for earlier years. Most of these revisions pertain to items in *Statistics of Income* tables which were not subsequently republished in tabular form even in issues of the primary source. Only by careful examination of the *text* in the successive issues of the primary source can we be sure of catching the revisions. A good secondary source *may* have caught these revisions, if it was not published before the revisions appeared in the primary source, but the chance is too slim to be relied upon.

Referring again to Table 14, we note that January, 1938, items are marked "preliminary." Many sources, even primary sources, do not systematically label preliminary items. But, where such a notation is encountered, it serves two main purposes. It notifies the user of the data that those particular figures are estimates, and not strictly comparable with the final figures for earlier dates. It also warns him that revisions are to be expected and prepares him to look for them in subsequent issues.

The most effective techniques for discovering revisions, in the use of sources, are discussed briefly in the notes on transcription methods which appear in Appendix B.

Recognition of a source as primary.—It is evidently necessary that the compiler of secondary data be able to discover which source is primary for any particular series. It is not always easy to answer this question, and not infrequently a single source is primary for certain series or tabulations and secondary for others. In the case of the economic treatise or textbook containing statistical material, little difficulty is likely to be encountered: ordinarily the author clearly states the sources of his secondary data and discusses in full the collection of such primary data as he presents. Likewise in the case of an economic or statistical memoir appearing in the journal of a learned society, the data presented are usually marked or described so clearly that they can be classified readily as primary or secondary. Similar remarks apply to the statistical publications of government agencies: these documents generally contain data assembled by the several publishing offices; and where secondary material is presented the original source is ordinarily stated.

The problem is somewhat more perplexing when the other prominent sources of economic data are considered. The official

[1] Washington, U. S. Treasury, 1933.

publications of trade associations and other journals devoted to the interests of particular industries give, with few exceptions, some data which are primary and some which are secondary; and in many instances the distinction between the two groups is not clear, or, if such distinction is made, reference to the original source of secondary figures is missing or inadequate. There are many nonspecialized periodicals, including the great weekly newspapers disseminating general financial, industrial, and commercial information, which present extensive tables of secondary data with relatively few indications of the original sources; but it will be found, although with difficulty in some cases, that several of these journals publish some series which are strictly primary. Numerous banks, large manufacturers, and public utility corporations now publish occasional or periodical statistical bulletins giving a review and prospect of business conditions, and here also the distinction between primary and secondary data is seldom clear. The practice of the more careful business forecasters is somewhat more satisfactory: in many of their publications adequate statement of the origin of data appears, and most of the data which they use are secondary.

If every secondary source systematically followed the rule of citing the reference to the primary source for each series of secondary data published, the difficulties of the user of the secondary source would largely disappear. He would assume that the source was primary for all its tabulations which did not bear such a reference citation. Although the practice of thus citing references is commendably general for all careful secondary sources, it is very far from universal. The investigator is therefore in frequent danger of finding, in some particular source, a series of data bearing no reference to another source, and of being in doubt whether the source in hand is truly primary for such series or has merely neglected to give reference citation.

He can sometimes dispel his uncertainty by careful examination of the source in hand. In the most favorable cases, the tabulation bears an unmistakable notation that the data are primary for the source. Even where no such notation appears, examination of the accompanying text may show beyond doubt that the source was in fact the original collecting agency. Sometimes, in the case of a periodical source, such textual evidence may be found more or less hidden in some earlier issue of the publication. Occasionally the investigator has no recourse except to communicate with the publisher and inquire specifically whether

the data in question are primary for that source. If all these devices fail, the investigator remains in doubt, and the only safe assumption to make is that the source is secondary for those data and that the primary source is "unknown."

As the statistician develops experience in the use of sources of economic data, he will come to have a broad acquaintance with them and to know which sources are primary for particular types of data. At the same time he will acquire skill in discovering the sorts of evidence which determine whether a source is primary or secondary. And he will build an increasing respect for those careful publications which always supply him with unambiguous information as to how they secured the data which they publish.

QUALITY OF SOURCES

In much of his statistical work the investigator will find it necessary to make a choice among available sources for any particular series of data. Even with respect to primary sources such a choice is frequently necessary, for often more than one source collects original data for a particular economic variable. For example: various agencies, official and otherwise, collect and publish data on unemployment in the United States; various financial newspapers give data, presumably original with them in each case, upon market interest rates; different trade journals in a particular industry give data, presumably originated by them, on prices and other aspects of the industry.

Where a choice must be made among primary sources, tests of reliability discussed below would normally be applied. But an additional test is usually helpful in selecting primary sources: the investigator should study the process of collection used by each such source and seek to determine which set of data more probably gives a true record of the economic variable under study. Where there is only one primary source the general rule obtains: unless there is very strong reason to the contrary, that primary source is preferred to all secondary sources.

In cases where secondary sources are used, the necessity of a choice among sources is much more common. Usually there are various secondary sources which publish the data of any particular series. Granting that the investigator has for good reason decided to use a secondary rather than a primary source, he is forced to choose an available secondary source. Unless he steadfastly resists the temptation, he is in danger of making this decision on the mere ground of ease. He will be tempted to "choose" the

secondary source which comes first and most readily to hand, or to take that source which arranges the data most nearly in the form desired for his own use and thus saves him the job of reorganizing them, or to take that source which includes all or most rather than just a fragment of the statistical information he needs.

Without denying the importance of making wise economies in all statistical work, we insist that the easy grabbing of data in the most convenient source is seldom a wise saving of labor. The statistician who desires to safeguard his analysis and results from imperfections entering at the very start—in his statistical raw materials—should rest his choice among sources upon a test of reliability rather than upon accessibility and convenience. Instead of being content with the first source he discovers, the statistician should discover several—perhaps all—of the existing sources and examine and compare them as to their quality. He may expect often to find that one source is preferable for part of his data and some other source for another part. Moreover, he may discover that the quality of a particular source changes over time: earlier issues of a publication may not be of the same quality as later issues, because of changes in control of the publication or for other reasons.

Evidences of quality.—No exhaustive list of the tests to be applied in determining the quality of a source can here be given. Some of the chief tests are mentioned briefly. The statistician, as he develops experience, will acquire a special skill in discovering evidences of quality. He should begin early in his statistical career to pause in his study of data and try to form a judgment of the quality of the source presenting those data, and this practice should become a settled habit in all his subsequent work. In dealing with sources the statistician needs to know as much as possible about their character, because the validity of his own work depends upon the reliability of the work done by his sources.

The major tests of the quality of a source have already been indicated or implied on previous pages. A secondary source which fails to give any citation of the primary source from which it quotes data is manifestly defective. A source which is found upon examination—not necessarily in connection with the series or tabulation immediately under study—to have made imperfect reproductions from a primary source is not entitled to confidence in the immediate case. Such imperfections may consist in actual errors in transcription, in faulty quotation of labels or units, in omission of significant footnotes, or in failure to reproduce all the

essential data. A source which presents data in poorly arranged and improperly organized tabulations, or otherwise uses slovenly methods, is to be suspected; but the use of workmanlike arrangements is not a sufficient test of quality. A source which contains internal evidence of conflict among data, whether in a single issue or in various issues (if it is a periodical source), is suspect: tables and groups of tables which are not consistent are evidence of poor quality in the source.

Purpose of publication.—General facts about the source, apart from facts related to the tabulations, sometimes aid in forming judgments as to its quality. The purpose of publication of a source often suggests question as to the reliability of its data, though knowledge of the purpose is seldom a sufficient basis for condemning a source. Sources published to promote sales, to advance the interests of an industrial or commercial or other group, to present the case of a political party, or to carry on any sort of propaganda, are suspect. Data published anonymously, or by an organization which is on the defensive, or under conditions which suggest a controversy, or in a form which reveals a strained attempt at "frankness," or to controvert inferences from other data, are generally suspect. To say that sources or data are suspect is not, of course, equivalent to condemnation; but the statistician has to protect himself by avoiding so far as possible use of data which are even likely to be unreliable.

A source which publishes data as a main or the chief function of its publication operation has ordinarily none of the foregoing counts, as to purpose, against it. If the publication of data is itself the purpose of the source, business reasons favor the maintenance of high quality. This is especially true if the source has been established over a long time and has an enviable reputation the maintenance of which is a business asset. Data compiled in a scientific inquiry, which bears the usual marks of careful scientific work, are usually of high quality, and so is the primary source presenting them. Data submitted under oath, or otherwise under conditions subjecting the issuer to a penalty, are generally of high quality. A source which is established and maintained by individuals requiring certain data, as a cooperative agency for collecting such data, is generally of high quality.

The daily newspaper, whether the nonspecialized paper which gives incidental attention to the presentation of statistical data or the financial or trade daily which regularly exhibits much of its news in numerical form, is on the whole the least satisfactory of

sources. There are numerous and notable exceptions to this statement, and the energy and persistence of certain newspapers have been chief contributors to the development of that broad foundation of economic and business data which now enables us to conduct empirical studies of economic phenomena. Nevertheless, one of the essential characteristics of competitive news publishing operates to minimize the efforts made to safeguard the accuracy of data printed in dailies. Speed is most important in the operation of a newspaper enterprise; and statistical material, like other "stories," must be issued while it is yet news. Moreover, once issued, it is in many cases beyond correction or revision: the interest in later and more accurate figures is frequently not considered sufficient to warrant expenditure of further time and space. In consequence, the student must have a discerning appreciation of the difficulties of the publisher if he would safely use newspaper data. Even after he has noted the wide differences in reliability as between different publishers, and also the considerable differences in reliability among the several groups or series of data in any one publication, he will need to be alert to detect the occasional serious slips in newspapers selected as most reliable.

COMPILATION FROM SEVERAL SOURCES

A practical problem of considerable difficulty and importance in the compilation of many series or bodies of data is the selection for joint use of a group of several sources. The inability to find in a single source all the data essential to a particular statistical investigation is perhaps the rule rather than the exception. Where primary sources are available for separate portions of the material, no problem ordinarily arises. When, however, some or all of the sources must be secondary, a decision is necessary as to which secondary sources should be used. The obvious plan is to choose the best, but in practice there are often too few which have the best qualifications or too many which appear equally good without being entirely satisfactory. It is very important that sufficient material be obtained, even at the expense of accepting some figures which are not quite so reliable as the best. Insufficient data sometimes preclude the finding of any worth-while results, whereas imperfect data often yield conclusions upon which one can generally rely within tolerable limits. This point is not, however, warrant for "solving" economic problems by using data which are seriously or generally defective. The investigator will ordinarily select the one best source for data of a particular sort;

but he will frequently need to set aside this rule, either because no one source gives a complete series of figures or because the best source is so unreliable that it must be checked by comparison with other sources on the chance that comparison even of poor sources may disclose errors.

Actual experience in the compilation of sections of the statistical material of a problem from various sources will arise frequently in the work of the statistician, and examples are therefore not given here. To illustrate the intricacy of piecing together sources, however, reference is made to a portion of a recent study of national income in the United States.[1] On pages 170 to 175 of the reference cited, descriptive comments indicate the sources used in compiling each of eight tables in the text of the document. Although many instances of joint use of several sources involve no such intricate compilations as those in the case cited, this example is by no means a rare exception.

Too great care can scarcely be given to the preliminary study leading to a selection of sources of data. A vigorous initial search to enlarge the field of choice and a painstaking examination of each possibility will give assurance that the data finally used are indeed the best available and will furnish the student with a knowledge of their imperfections and the consequent limitations upon the results of their analysis. Failure to make a thorough initial survey may result in the compilation of faulty data which must be abandoned later when better sources are discovered. Frequently the waste occasioned by such a blunder is very great, for it includes not only the labor of securing the improper data but also the work incident to such analyses as may be applied to those data before they are finally discarded.

---

[1] "National Income, 1929–1932," Senate Document 124, 73rd Congress, 2nd session, Washington, 1934

CHAPTER IV

COLLECTION OF PRIMARY DATA

PLANNING A STATISTICAL INVESTIGATION

In general, investigations in economic statistics can be conducted by the use of available secondary data; but important instances may arise in which existing data are inadequate for thorough study, and some or all material must be secured by an original *survey*. Moreover, even in the ordinary case in which he uses secondary data (above, page 39), the investigator needs to be familiar with the principles governing the collection process if he is to handle intelligently data secured from available sources. Although certain general principles pertain to the collection of primary data, no set of rules of procedure which is applicable to any considerable number of the widely differing cases which confront the statistical practitioner can be constructed.

The situation which generally arises in practice may be described as follows. The formulation of a particular problem in economic statistics consists in propounding certain questions which are to be answered by a statistical inquiry; for, although actual study of available or newly collected figures frequently leads to the establishment of propositions which were not originally suggested, such study is normally directed to the testing of specific hypotheses rather than to the blind search for hidden truths. Such a blind search—a sort of statistical fishing expedition—is indeed sometimes made: a body of data is examined in the hope that it will suggest questions and their answers.

The investigator foresees that the answering of the several questions proposed necessitates the preparation of certain summary tables (see below, Chap. VI), the construction of specific charts, or the calculation of particular statistical ratios, index numbers, averages, or other derived statistical numbers. As soon as these requisites are listed, precise lists of the series of data which will be needed can be made, and consequently general tables which would contain the material essential to the study can be described. In more technical terms, this description consists of defining the variables to be measured by statistics, and hence of

specifying labels and units of statistical items. Thorough canvass of available secondary material is the next step in the planning process. This enables the student to determine whether there are published data for the purpose in view, and whether they are sufficiently comprehensive and detailed and satisfactorily reliable.

In so far as published data are inadequate, the lack must be supplied by primary figures. At this stage of his research, therefore, the investigator will have in mind the outline of those general tables, if any, the data for which must be obtained by an original survey. The information which appears necessary for a solution of the problem may in fact not be completely obtainable because of obstacles in the way of effective collection. Nevertheless, the logical way to attack the problem is that outlined above, and the precise forms of general tables sought must largely control both in general and in detail the gathering of any new data.

THE PRIMARY SURVEY AND THE INFORMANTS IT REACHES

Once a decision is reached concerning the specific primary data desired, the sources ("sources" does not here refer, of course, to sources of published data) from which to secure the information must be discovered. Presumably the facts are in the possession of numerous persons or corporations or other associations. If the possible informants are very numerous, the question at once arises as to whether or not all shall be addressed or interviewed; the answer will probably hinge upon the time and money available for the survey. With few exceptions, statistical surveys are expensive undertakings; the time which elapses and the expenses which are incurred, from the preparation of the initial inquiry blank to the completion of the desired general tables, frequently far exceed all estimates. Therefore, unless the particular problem in view is exceedingly important, the student will be wise to refrain from an attempt to make an exhaustive examination of every possible source of information.

Assuming that the prospective informants are numerous and that all cannot be interviewed, it is necessary to decide how many and which informants shall be addressed. The determination of the extent of the survey—the portion of the total number of possible informants to be addressed—involves considerations of *sampling* technique and theory. Brief comment upon this highly important topic appears below (page 216), but the nature of the issue involved should be understood in connection with the planning of a survey. Whenever statistical data pertaining to a limited

group of cases within a larger and more comprehensive list of cases are used as the basis for inferences concerning the larger list, the limited group is called the *sample* and the larger list is called the *population*. The word *population* is here used, of course, more generally than in common speech—it can refer to a list of cases of any sort, inanimate or institutional, as well as individual human beings. Appropriate methods for selecting a sample from a population, and the theoretical limitations upon inferences concerning the population but derived from the sample, constitute one of the most important and difficult branches of statistics (see below, page 218). The use of a limited list of informants in making a primary survey is one somewhat special case of sampling.

The determination of the extent of the survey is closely related also to the decision as to the manner of conducting the inquiry. Some methods of collecting data are much less effective than others and must therefore cover many more informants in order to secure data of equal volume or reliability. Several considerations govern the decision of these two questions, how many to address and how to address them. The importance of the investigation, and the time and money available for its conduct, have already been noted as controlling elements in the case. Quite as important as the number of informants is the degree of their intelligence: the results of the survey will differ greatly according as the questions are addressed to a selected group of exceptionally competent persons or to a heterogeneous list of people of all grades of ability and training. Moreover, the specialized knowledge of the individual informant, as well as his general intelligence, should be considered: a person whose experience or interests especially qualify him to give the desired information is presumably much more valuable as a witness than the individual having only fragmentary or incidental knowledge of the facts.

An important factor also is the attitude of the prospective informant toward the ultimate objects of the study: his active interest in the anticipated results of the survey will ensure his responsiveness and the accuracy of his replies. If he is compelled by law to give information, this point is less important—although even the census, which is compulsory, is careful to seek only facts which are of quite general interest and to avoid asking questions which might arouse resentment. If, however, the giving of information is not mandatory, as is usually the case in an economic survey, the interest of the individual addressed must be aroused and he must be convinced that the outcome of the study will

certainly not injure him and that it will probably benefit him. The answering of even relatively few questions with the care and accuracy essential to the statistician involves the expenditure of much time and effort on the part of the informant and frequently also invites him to disclose facts concerning his affairs which he has long preferred to keep secret. Courtesy is an absolutely necessary qualification of the field worker in statistics, and he should also have a fitting appreciation of the actual extent of the favors he frequently asks.

METHODS OF CONDUCTING THE SURVEY

The two important methods of getting new statistical data are by *enumeration* and by *registration*. We are concerned here chiefly with the former. Enumeration is the isolated collection of data for a specific purpose, whereas registration is the continuous and automatic return of a report of each fact that pertains to a given statistical problem, either as soon as the fact becomes known or at specified regular intervals. Thus the decennial census is an enumeration, but the annual filing of Federal income tax returns by individuals is registration; the official postcard collection of unemployment data in November, 1937, was an enumeration, but the monthly reporting of a railway to the Interstate Commerce Commission is registration. Registration may consist in the rendering of continuous or periodical returns by an informant possessing the desired facts, or in the regular submission of data by reporters actually associated by more or less direct ties with a central statistical office to which the reports are directed. Submission of bank statements to the Comptroller of the Currency, reporting of bank clearings to the *Commercial and Financial Chronicle*, union secretaries' returns on labor conditions to a state industrial board, and filing of individual income tax returns are instances of registration of the first sort. The returns to the Department of Agriculture by its crop-reporting service, the rendering of information on credit conditions to a mercantile agency such as Dun & Bradstreet, and the reporting of local prices of a specific commodity by representatives of a trade journal, are examples of the second type of registration. This second plan of securing current data is very widely used; and it has many advantages, for the reporters are generally trained and experienced in securing the special information which they seek.

Methods of enumeration.—For the student attacking a new problem, however, in which entirely new statistical material must be

secured, registration is not available; some scheme of enumeration is needed. The three chief plans of conducting an enumeration are: a personal survey by the investigator, the interviewing of informants by enumerators, the mailing of blanks to the informants. The first plan is available only for very small studies in which one individual can hope to cover the ground in a reasonably short time; but, for such problems, it is an excellent scheme. Its results, where it is applicable, are likely to be notably superior to those obtainable by any of the other methods. The fact that a single individual, who assembles all the data, is able to decide according to his single judgment and therefore upon nearly uniform rules all perplexing borderline cases ensures the compilation of more homogeneous statistical material than can be secured by any other method.

In order, however, to reach any considerable number of informants in a time which is only moderately long, one of the other methods of enumeration must be used. From the point of view of the accuracy and effectiveness of an intensive survey covering a large list of informants, the method of enumerators has marked advantages. It is the plan followed by the United States Census, and is widely used in other studies of sufficient importance to warrant the necessary expenditure. According to this scheme, the blank forms, upon which data are to be entered, are placed in the hands of reporters who have received instruction as to the proper manner of entering the items; and these reporters interview the individual informants and fill in the blanks in their presence or after collecting facts from each of them. In this manner there is an approach to completeness in covering the field; the returns are sent in promptly; and, provided the reporters are competent and well trained, there is a high degree of uniformity in the answering of questions raised by exceptional cases. Moreover, answers to a longer and more complicated list of questions can be secured in this manner than by the direct-mail scheme.

The other plan—mailing of blanks to informants—is most often used when the field to be covered is fairly large and the time and money to be used in the study are less than might be desired. The mailing of forms enables the investigator to address a very large number of possible informants, but the returns are usually far less numerous than the forms sent out: in most cases only from 10 to 40 per cent of the first letters bring replies. Also, frequently the individuals who reply comprise only the more intelligent or better-informed among those addressed; and the sample of actual

returns may therefore be biased. Moreover, the saving of time over the method of enumerators is partly illusory because of the long delay incident to the receipt of replies and the handling of the large follow-up correspondence occasioned by fragmentary and incorrect answers.

THE BLANK FORM

When decisions have been reached as to the specific information to be sought, the individuals to be addressed, and the method to be used in collecting the data, the investigator is in a position to prepare the blank forms. Each of these three decisions has an important bearing upon the form and complexity of the information blank. Although a small survey may be conducted without any blank form, such procedure is seldom desirable; and, with almost no exceptions, the returns in a survey of any considerable magnitude should be made upon sheets or cards having uniform rulings and uniform labels and other notations, Here the influence of the desired general table (see below, page 60) is felt; the form of that table and the technique of its compilation from the returns control in a measure the nature and arrangement of the questions on the blank. On the other hand, practical points pertaining to the securing of answers to the questions are the chief guides in framing the blank; the form of the blank is decisively controlled by considerations relative to obtaining the data rather than utilizing the results.

The conflict, between considerations pertaining to the securing of information and those pertaining to tabulating the information secured, can generally be resolved by careful study in preparing the forms. It is in the highest degree unwise to prepare a blank, except in an extremely simple type of survey which seeks data on only one or at most very few variables, without looking ahead to the tabulation problem. Nevertheless, the great advantage of simplicity and logical arrangement of the form, in promoting cooperation by the informants, frequently warrants subordinating considerations of ease in tabulating. The investigator must aim at ease and economy for his informants rather than for himself and his technical assistants.

Numerous instances arise in which a form serves more than one purpose, and here the framer of the form must study not only considerations relative to responsiveness of informants and to facility in tabulation, but considerations relative to the effective use of the returns for the additional purpo

the method of registration rather than that of enumeration, consider the Federal income tax blank for individuals. The primary purpose of the blank is to supply the Treasury with information concerning the tax liability of each individual; the tabulation of the data, for the use of Congress and other agencies or individuals interested, is only a secondary purpose. In framing the blank form, therefore, three major sets of considerations are studied: (1) securing all the facts needed, in a form facilitating their effective use by the Treasury in checking individual tax liabilities; (2) easing the task of the taxpayer in supplying the information fully and accurately; and (3) rendering speedy and economical the preparation of the desired tables. Other instances, in which more or less conflicting considerations must be balanced in framing a blank, will be encountered by the student in his study of surveys actually made or proposed.

Framing of questions.—In regard to the nature of the questions, Professor Bowley says:

> The questions must be so clear that a misunderstanding is impossible, and so framed that the answers will be perfectly definite, such as a simple number, or "yes" or "no." They must be such as cannot give offense, or appear inquisitorial, or lead to partisan answers, or suppression of part of the facts. The mean must be found between asking more than will be readily answered and less than is wanted for the purpose in hand. The forms must contain necessary instructions, making mistakes difficult, but must not be too complex. The exact degree of accuracy required, whether the answers are to be correct to shillings or pence, to months or days, must be decided.[1]

The precise meaning of every term used in the blank must be clear, and frequently questions must be amplified by footnotes because of the possibility that exceptional cases will be reported incorrectly. Sometimes these notes are so numerous and involved that a set of instructions must accompany the blank form. Usually the informant should not be asked to calculate and fill in totals or averages or ratios derived from figures which he has given, although such derived entries sometimes serve as checks upon the accuracy of his replies. Ordinarily some questions can be introduced which provide partial or complete confirmation of the answers to others, for instance, age as well as date of birth, quantity as well as value of product, interest paid as well as amount of debt, capacity as well as output of plant, places and dates as well as duration of service. The way in which the returns are to be collected

[1] "Elements of Statistics," P. S. King & Son, London, 1920, p. 15.

will have an important bearing on the number and form of the questions asked. Thus a form sent through the mail must ordinarily have fewer and simpler questions than one filled out in the presence of or by an enumerator.

EDITING THE RETURNS

As the survey progresses, the blanks (filled in either by the informants or by enumerators in the field) are returned to the central office where the information they yield is to be tabulated and used. Before any satisfactory tabulation is possible, each individual return must be examined in detail to ascertain whether or not it has been answered in full and, so far as internal evidence shows, accurately. This work of *editing* requires skill and scientific impartiality to a very high degree; for, although it involves making additions to or even changes in the answers in some instances, it must be done in such way that there is no possible falsification of the original return. Bailey and Cummings name four types of editing: editing for consistency, uniformity, completeness, and accuracy.[1]

Comparison of the answers to those questions on the blank which are designed to be mutually confirmatory indicates, without proving conclusively, whether there is consistency in the return. If the answers to two such questions appear to be mutually contradictory, it is necessary to determine which, if either, is correct. The most obvious procedure is to address another inquiry to the informant for the purpose of checking on the doubtful question; but this supplementary inquiry will not always yield a decisive result, and it will in any case consume considerable time and effort. Moreover, the conflicting answers on the return may incline the investigator to question the veracity of the informant, or at least to doubt his knowledge of the facts; and, in such instances, the particular return will probably be discarded entirely. Even if there are no adequate grounds for discarding the entire return, the inability to explain and remove the contradiction between the two answers may necessitate the omission of these particular answers on this return from the final tabulation of returns. The investigator must guard against too hasty decisions to throw out doubtful replies, if he would avoid losing important items of information bearing upon his problem; and he must

[1] The discussion in the following paragraphs is based largely upon the treatment by W. B. Bailey and John Cummings, in their "Statistics," A. C. McClurg and Company, Chicago, 1917.

practice careful restraint against allowing such decisions to be reached because some item or items on particular returns "look extreme" or are unexpected. Otherwise, his data are spoiled by a subjective element, and his survey loses its scientific integrity.

Lack of uniformity in filling in the various returns in the survey must ordinarily be charged against the blank form itself, on the ground that it was not designed to preclude misinterpretation, or against inadequate skill or incomplete training of the enumerators. Nevertheless, even in primary collections conducted with the utmost care, there will be an occasional return in which the replies are submitted in the wrong manner. Perhaps the most frequent mistake of this sort is in the statement of units: time may be expressed in years rather than months, weight in tons rather than pounds, value in dollars rather than thousands of dollars. Such instances of reporting data in some unit other than that specified or desired in the blank are usually recognized easily in the returns; however, the possibility that they will not be recognized in the editing process is sufficiently great to justify extreme care in drafting the blank so that they cannot occur. When they are recognized, further correspondence with the informant may be needed to clear away the difficulty; although in some cases the error is so obvious that the editor can remedy it without the slightest fear of mistake.

In addition to the liability to misuse of units, other less apparent and fortunately less frequent causes of nonuniformity exist; for instance, the stating of income for a fiscal year when it is sought for the calendar year, the reporting of a price for the end of a month when it should be the average price for the whole month, the return of the rate of turnover in labor based upon the maximum payroll when it is sought on the basis of the average payroll. These mistakes are usually difficult to discover from internal evidence and must therefore be the more surely guarded against in the blank itself; but when they are discovered during the editing they should be adjusted by making obvious corrections or, as is more frequently the case, by further correspondence.

Editing for completeness is, in the main, a straightforward operation. Certain spaces often appear on the blank form, such as totals and percentage ratios, which the informant was not intended to fill in. The office worker, in editing each return, calculates and enters these items as a matter of course. Furthermore, he probably has certain distinctive symbols connected with the tabulating process which he enters automatically as he edits—

he *codes* the return. The only instance of incompleteness in the return which is likely to lead to further correspondence is the actual omission of the reply to a specific question on the blank. In such a case an answer must be obtained by subsequent inquiry unless that question is to be ignored, as far as the particular return is concerned, in tabulation.

Editing for accuracy is the most delicate of all, and can rarely be pushed very far. Inaccuracies, other than inconsistencies, are seldom apparent from internal evidence in the return. The skillful editor develops through wide experience a special aptitude for discovering such errors; the novice must not be discouraged if he misses most of them. A glaring error is always found easily, but the less striking mistake passes unnoticed and yet may work serious damage to the statistical study. When such mistakes are discovered by the editor, he is scarcely ever justified in correcting them without further recourse to the informant; for the mistake is usually of such sort that its existence but not its size can be determined by inspection of the return.

The editor should bear in mind certain general rules in carrying on his work. He must never destroy, by clipping or erasure, the original reply, but must enter any corrected items separately, preferably in a distinctive ink. Indeed, all the marks which he makes upon the return should be clearly distinguishable from the original—a subsequent investigator should be able to identify the information exactly as supplied by the informant. The attitude of mind of the editor must be such that he regards the original return as evidence which must in no way be lost or altered. If changes appear necessary in certain entries before tabulation, they may be added; but the primary replies must always be available for further reference and study. The process of editing is by no means an unimportant and routine operation; rather it requires marked ability, scrupulous care, and a rigid adherence to scientific objectivity.

CHAPTER V

CONSTRUCTION OF GENERAL TABLES

GENERAL AND SUMMARY TABLES

The process of compiling secondary data for use in an analysis normally consists in abstracting the desired items or series from published tables. From the point of view of form and content there are two broad classes of tables, general and summary. The *general table* is designed to give a detailed presentation of the items furnished by the primary collection, completely and accurately, and without any purpose of suggesting a particular interpretation by arranging the items in a particular way or calculating from them particular derived numbers. The general table is, then, merely a repository of numerical facts, and it should be as compact as accessibility of all the items will allow. One of the obstacles to the publication of truly general tables is in fact the difficulty of securing compactness: one has merely to glance at some of the general tables published in the quarto volumes of a decennial census to be impressed with the wasteful use of vacant space in the body of such tables.

A completely general table would, of course, list categorically each item of information contained in each separate return. Thus, suppose a survey of employable unemployed labor in a community had secured 840 returns, each showing the age of a workman, his occupation, the industry in which he was formerly employed, the hourly rate of wages at which he was employed, the number of weeks he has been unemployed, and the number of dependents in his family. Suppose that each return bears a serial number, to avoid disclosing the identity of the workman. A completely general table would then include seven columns—one for the serial number and one for entering specifically each of the six facts reported—and 840 rows. It would, in other words, be merely a convenient tabular organization of *all* the facts presented on every return. Obviously, such complete generality of tabulation is practically impossible, because of the expense of space, in all except the simplest and most limited surveys. For any body of data which is at all voluminous, the returns have to be sorted into

classes. The class figures, rather than the detailed facts of each separate return, appear in the table. Such a table is regarded as general, in the sense defined above.

The *summary table* is a condensed presentation in which only a portion of the original material appears, or in which the items are derived by computation from the original material, or in which the arrangement is consciously designed to suggest or facilitate a particular interpretation of the items. Thus the summary table may be regarded as prepared from a general table, real or hypothetical, by a process of abstracting some of the items of the general table, or basing certain computations on those items, or arranging those items with a view to emphasizing certain comparisons. The line between the two classes is not clearly drawn, and there are tables which should strictly be called "summary" but which contain so large a portion of the original detail that they are accepted as general tables. Moreover, in some instances, data collected in a survey are entered directly into summary tables, without appearing in a general table. But ordinarily a primary source which does not present general tables is not entirely satisfactory.

Two main facts favor the use of general tables as sources: They furnish complete detail and thus enable the student to be confident that important figures have not been concealed by a process of summarizing, and they are ordinarily more reliable numerically than the summary tables which have been obtained from them by processes of transcription and computation. On the other hand, in many investigations interest centers only on a limited portion of the evidence furnished by statistics, and the data in a properly constructed summary table can be accepted as sufficient. In such cases the controlling consideration is the saving of time which attends the use of the condensed material in the summary table; whether the loss of time involved in using general tables as sources is justified by the advantages of completeness and assured accuracy afforded by such sources must be decided in each specific problem.

THE CLASSIFICATION SCHEME IN A GENERAL TABLE

In taking up the considerations bearing upon the problem of tabulation we shall speak first of general tables. It is assumed that the data of a statistical investigation have been assembled by means of a primary survey and that the returns have been edited.

The next step is the orderly presentation of the data in tabular form. An operation involved in this step is the classification and counting of the material. Thus, if the survey pertains to weekly wages, the returns probably give not only the dollar amount of wages but also specify whether the wages in each individual case were received by a man or woman, whether in a closed or open shop, whether in a textile or steel or leather or paper or other specific plant, in which of various specific occupations, in which of several definitely known cities; and indeed, if the survey is at all extensive and thorough, the results are likely to contain information which renders possible classification according to numerous characteristics. If the survey was properly planned, all or most of these possible bases of classification should have been foreseen before the collection of data began.

Each of the several bases of classification implies a particular rule of classification. It might be supposed that some freedom of selection would exist in regard to the rule of classification for any one classification scheme; thus, for a distribution on a geographical basis, the returns might be grouped according to cities or counties or states. For the preparation of a *general* table, however, most of these alternatives disappear; and, in the detailed presentation of the complete data, the choice of classification rule is quite narrowly restricted for any particular basis of classification. For example, in the above case, classification according to cities is clearly desirable; for the grouping of cities in counties or states involves a summarization of the data. Moreover, complete freedom in the order of arrangement of the cities does not exist; for some arrangements would suggest particular comparisons of the items and thus would evidently lead to a summary table, in the sense indicated by the definition.

Obviously, in many cases, the strict requirements implied by the definition of *general table* cannot be realized, particularly because of the physical obstacles to handling and presenting great masses of data. Seldom, even under ideal conditions, can a general table actually present *all* the details of statistical information afforded by the individual returns collected in a survey. Suppose, for instance, that the basis of distribution in the above illustration is to be according to the amount of wages received. Strictly, separate entries should be made for all the different wage amounts received—there should be a separate group in the table for every different wage. If wages are stated in cents and there is a wide variation among the returns of the survey, such

detailed presentation may prove physically impossible. This then suggests some of the reasons for relaxing the definition of *general table* in many instances—*some* summarization of the original returns is often inevitable if any presentation is practically to be made. In thus presenting, as a general table, the result of a tabulation which involves concealing some of the detailed information given by the original returns, the investigator must take care to keep summarization to a minimum and to select groupings and arrangements which do not suggest inferences tending to falsify the facts and which do not restrict the use of the tabulated information by other investigators.

Specification of groups.—As soon as the rule of classification is determined, each return can be assigned to its proper group. Each class in the rule must be specified exactly: the rule of classification must be unambiguous. Otherwise, borderline cases may give serious trouble. Although these borderline cases arise most conspicuously in frequency series—in classification according to size of a variable (such as wages)—they can appear in the formation of time or categorical series. In all types of series the classes in the rule of classification must be mutually exclusive—overlapping of classes must be avoided. Whereas actual experience in the classification of data will most emphatically show how frequent and diverse the borderline cases are, brief reflection should suffice to suggest to the reader some of the chief possibilities. These difficulties can usually be removed effectively by proper specification of the classification groups, but much trouble can be avoided by the exercise of foresight in planning the collection of data. Otherwise items may be reported in the returns in a manner which prevents their precise tabulation.

Certain special requisites for satisfactory groups apply to frequency series. It is highly desirable that all the intervals be of equal width and that there be no groups of the form "all under". and "all over." Tables 15 and 16 are instances in which these rules have been, respectively, ignored and observed. Moreover, known tendencies for unduly large concentrations of returns at certain points should be allowed for by a grouping which fairly distributes these points among the several intervals. Thus (see Table 17) at least for older persons there is a marked tendency for ages to be stated in multiples of 10, and a considerable tendency also to concentrate in multiples of 5. If these age data are to be presented in groups 10 years wide, a selection of intervals is desirable which places a multiple of 10 near the center of an

TABLE 15

FREQUENCY DISTRIBUTION OF EMPLOYEES, ACCORDING TO EARNINGS PER HOUR, IN THE BOOT AND SHOE INDUSTRY (FITTING OR STITCHING DEPARTMENT, VAMPERS, MALE) IN 1932*

Earnings[a]	Number of employees	Earnings[a]	Number of employees
Under 12	1	60–70	53
12–16	2	70–80	34
16–20	1	80–90	19
20–25	2	90–100	7
25–30	8	100–120	2
30–40	45	120–140	1
40–50	47	140 and over	1
50–60	60		

* Unit, cents per hour. Source: "Wages and Hours of Labor in the Boot and Shoe Industry, 1910 to 1932," Washington, U. S. Bureau of Labor Statistics, Bulletin 579, March, 1933, p. 82.
[a] Lower limit inclusive.

TABLE 16

FREQUENCY DISTRIBUTION OF EMPLOYEES, ACCORDING TO WAGES PER HOUR, IN COTTON MILLS, IN 1900*

Wages[a]	Number of employees	Wages[a]	Number of employees
5–5.9	6	9–9 9	18
6–6 9	7	10–10 9	18
7–7.9	3	11–11.9	2
8–8.9	11	12–12 9	3

* Data for "roving-frame tenders, southern states, males" Source: "Twelfth Census of the United States, 1900, Special Report on Employees and Wages," Washington, U. S. Census Office, 1903, p. 42.
[a] Unit: cents per hour.

interval, although this involves placing a multiple of 5 near the end.[1]

The counting process.—A device often used in carrying data from the returns to the general table is the individual card. The area of such a card is divided into sections, each reserved for the answer of a particular question which may occur in a particular return. The fact given in each return, in answer to each question, is entered on the card by making an appropriate mark (for example, by punching a hole in the proper space of the card). The entry on the card may merely be a mark indicating that the particular return shows the quality in question—for example, the mark may

[1] For methods of adjustment in such cases, the student is referred to some text on population statistics, or statistical graduation.

TABLE 17

DISTRIBUTION OF THE POPULATION OF THE UNITED STATES IN 1930, ACCORDING TO AGE*

Age (years)	Number	Age (years)	Number	Age (years)	Number
Under 1	2,191	35	2,051	70	561
1	2,165	36	1,800	71	347
2	2,326	37	1,709	72	402
3	2,394	38	1,963	73	338
4	2,369	39	1,685	74	302
5	2,505	40	2,090	75	323
6	2,515	41	1,327	76	242
7	2,470	42	1,756	77	194
8	2,604	43	1,434	78	188
9	2,513	44	1,382	79	160
10	2,501	45	1,727	80	170
11	2,319	46	1,331	81	104
12	2,480	47	1,278	82	102
13	2,322	48	1,433	83	85
14	2,382	49	1,273	84	74
15	2,296	50	1,638	85	66
16	2,367	51	951	86	49
17	2,296	52	1,237	87	38
18	2,358	53	1,061	88	29
19	2,235	54	1,088	89	23
20	2,222	55	1,100	90	21
21	2,211	56	959	91	10
22	2,203	57	861	92	9
23	2,131	58	917	93	6
24	2,104	59	809	94	5
25	2,086	60	1,075	95	4
26	1,986	61	605	96	3
27	1,882	62	743	97	2
28	1,959	63	703	98	2
29	1,920	64	626	99	1
30	2,197	65	745	100 and over	4
31	1,549	66	491	Unknown	94
32	1,875	67	500		
33	1,731	68	544		
34	1,768	69	490	Total	122,775

* Numbers stated in thousands. Source "Fifteenth Census of the United States, 1930, Population, Vol II," Washington, U. S. Department of Commerce, 1933, pp 593–594.

show that the return was filed by a molder, in a survey of steel workers. Or, the entry on the card may be a number—for example, the amount of gross sales, in a survey of department stores.

The cards prepared in this way are then ready for tabulation by hand or by machine. The hand method consists in sorting the

cards into piles according to the rule of classification, and is satisfactorily effective only when the returns contain replies to but one or at most very few questions. The counting operation follows. If the entries are mere marks indicating the existence of a quality, counting consists merely in finding how many cards are in each pile. If the entries are statistical numbers, counting consists in adding up the numbers on the cards in each pile. With respect to numerical entries, of course, the rule of classification being followed may be according to size. Then each card is sorted into the size class to which it belongs, and the cases in each such class are merely counted to get the class frequency.

In the machine method the cards are automatically sorted and counted by an electrical device. The basic principle is that the machine registers whenever a hole in a particular card allows an electrical contact to be made in the circuit belonging to the group being counted. The machine method is adaptable to both cases—when entries are mere indications of existence of a quality and when they are numbers. The electrical machine is much more efficient and reliable than the hand method, but it is expensive and scarcely available for any problems in which the volume of work to be done is only moderately large.

In handling classification and tabulation tasks by hand methods, and indeed in all practical statistical work of a routine sort, the student should ensure the efficiency and accuracy of his performance by every possible means. The principal ways in which he can accomplish this are: by devising and regularly following in routine work an orderly and logical plan, as concerns both the broad outline and the minor details of the operation; by providing at every stage of the process for as many automatic checks as possible on the accuracy of method and results; and by arranging for an independent verification of the entire task by another person or (though this is less satisfactory) by himself at another time. The beginner will be disposed to regard some of these safeguards as needlessly boresome, but experience in the use of faulty data will soon convince him that every effective precaution is worth while.

ORDERS OF TABULATION

When the classification is according to one variable or quality only—for example, according to the amount of wage or the place of employment—the resulting table is a simple structure with a single rule of classification and the corresponding items. Such tables are

said to be of the *first order;* Tables 1 to 17 furnish illustrations of this type. Generally classification is also possible according to one or more additional coordinate characteristics (variables or qualities), and each main group can in many cases be subclassified according to several subordinate characteristics. If the process is carried to the limit, the multiple classification can be made exhaustive and the material furnished by the returns can be digested completely. If the classification desired is thus exhaustive, or if it is complicated without being truly exhaustive, the crude methods of classifying and counting discussed above will not be practicable. It will be necessary to establish a systematic tabulation process and, if the volume of data is very large, to provide for sorting and counting by machine. In any case, the various cards need to be sorted several times: first according to one rule of classification, then each of these classes according to another rule of classification, and so on.

The simplest type of tabulation, in which the classification is according to a single attribute (quality) or variable, leads to a table of the first order, such as those shown above. A table in which the classification depends upon two independent coordinate attributes or variables (as in Table 18), or in which there is a

TABLE 18

PER CENT OF EMPLOYEES ("CUTTERS, VAMP AND WHOLE SHOE, HAND, MALE") IN THE BOOT AND SHOE INDUSTRY HAVING SPECIFIED EARNINGS PER HOUR*

Year	Under 40 cents	40 cents, but under 70 cents	70 cents or over
1926	2	31	68
1928	2	28	71
1930	3	30	67
1932	9	49	40

* Source: "Wages and Hours of Labor in the Boot and Shoe Industry, 1910 to 1932," Washington, U. S. Bureau of Labor Statistics, Bulletin 579, March, 1933, p. 24.

subclassification within all or several of the main groups according to a single subordinate attribute or variable (as in Table 19), is a table of the *second order.*[1] In Table 19 the classification according to time is subordinate to that according to amount of wage, whereas no distinction is made in Table 18. Thus, even in cases

[1] Neither of these tables is strictly general: the calculation of percentages has concealed the actual numbers of workers covered. The "summarization" here is, however, not a serious obstacle to treating these tables as general.

as moderately complicated as these the arrangement must be governed by the importance of the facts to be presented.

Tables of order higher than the second are very widely useful in presenting statistical material, especially in general tables, but if the multiple classification is pushed too far the complexity of the table becomes so great that the reader can only with difficulty comprehend the significance of the grouping or conveniently use

TABLE 19

PER CENT OF EMPLOYEES ("CUTTERS, VAMP AND WHOLE SHOE, HAND, MALE") IN THE BOOT AND SHOE INDUSTRY HAVING SPECIFIED EARNINGS PER HOUR*

Earnings	Per cent
Under 40 cents	
1926	2
1928	2
1930	3
1932	9
40 cents, but under 70	
1926	31
1928	28
1930	30
1932	49
70 cents, or over	
1926	68
1928	71
1930	67
1932	40

* Source: see Table 18.

TABLE 20

RELATIVE HOURLY EARNINGS OF WEAVERS IN THE COTTON GOODS AND WOOLEN AND WORSTED GOODS MANUFACTURING INDUSTRIES OF THE UNITED STATES

Year	Cotton		Woolen and worsted	
	Male	Female	Male	Female
1920	337	322	348	379
1922	229	232	266	292
1924	264	262	302	332
1926	233	229	281	305
1928	231	226	284	307
1930	235	232	274	294

* Unit: per cent of the 1913 average. Source: "Wages and Hours of Labor in Cotton-Goods Manufacturing, 1910 to 1930," p. 7, and "Wages and Hours of Labor in Woolen and Worsted Goods Manufacturing, 1932," p. 8, Washington, U. S. Bureau of Labor Statistics, Bulletin 539, June, 1931, and Bulletin 584, July, 1933.

the data in further statistical work. In fact, construction of tables of order greater than the fourth is seldom wise, for even this usually involves two main coordinate classifications, each with its two subordinate groupings. Tables 20 and 21 are instances of tabulations of the third and fourth orders, respectively. Where a table contains a subordinate grouping, the arrangement can be altered so that the property regarded as subordinate becomes primary. Table 22 presents the material of Table 20 in an altered form, with the result that the main grouping in columns is according to male and female operatives instead of according to cotton and woolen mills.

TABLE 21

NUMBER OF COMMON STOCKHOLDERS IN SELECTED LARGE CORPORATIONS AT END OF EACH QUARTER IN 1937*

Quarter	Actual number (*in thousands*)						Percentage of 1929 average[a]					
	Pennsylvania Railroad Co.		U. S. Steel Corporation		American Telephone and Telegraph Co.		Pennsylvania Railroad Co.		U. S. Steel Corporation		American Telephone and Telegraph Co.	
	Domestic	Foreign	Domestic	Foreign	Domestic	Foreign	Domestic	Foreign	Domestic	Foreign	Domestic	Foreign
I	214	3	161	3	632	7	125	100	150	150	139	140
II	212	3	158	3	631	7	124	100	148	150	139	140
III	212	3	156	3	631	7	124	100	146	150	139	140
IV	213	3	161	3	634	7	125	100	150	150	140	140

* Source. *Survey of Current Business*, February, 1938, p. 36.
[a] Computation based on data in columns 2–7 and 1929 average in *Survey of Current Business*, 1936 Supplement, p. 64.

TABLE 22

RELATIVE HOURLY EARNINGS OF WEAVERS IN THE COTTON GOODS AND WOOLEN AND WORSTED GOODS MANUFACTURING INDUSTRIES OF THE UNITED STATES*

Year	Male		Female	
	Cotton	Woolen and worsted	Cotton	Woolen and worsted
1920	337	348	322	379
1922	229	266	232	292
1924	264	302	262	332
1926	233	281	229	305
1928	231	284	226	307
1930	235	274	232	294

* Unit: per cent of the 1913 average. Source: see Table 20.

The arbitrary placing of emphasis on certain classifications by determining which shall be leading and which subordinate implies the formation of a summary rather than a general table. Even in a general table, however, equal prominence can frequently not be given to all the grouping schemes. From the point of view of the general table, which aims merely to present data, this question of arrangement can be settled solely by consideration of the relative importance of the two bases of grouping; but, in a summary table, the arrangement is controlled by the necessity of placing adjacent to one another those columns between which the chief comparisons are sought. The line between general and summary tables is often so difficult to draw that this principle of arrangement finds some place in the planning of general tables.

PRACTICAL DETAILS IN TABLE CONSTRUCTION

In addition to the group arrangement of the table, certain details of form merit careful attention; and these apply to summary as well as to general tables. The titles of the columns are called *captions* and those of the rows *stubs*. The order of entry of the captions from left to right, and of the stubs from the top downward, is of moment. The most important positions in the table are the left column and the top row, and the advantages of position decrease gradually toward the right and toward the bottom, except that the right column and bottom row have some preference over their immediate neighbors. Formerly totals were customarily placed at the right and bottom; but the practice of placing them at the left and top is now fairly general, and these two principal positions in the table are often reserved for totals of rows and columns, respectively. For the arrangement of the other items, the order of rank remains from left to right and from top downward.

In a distribution according to size (a frequency series) or according to time (a time series), the order of arrangement presents few difficulties; consecutive rows or columns belong to consecutive points or intervals of size or time. The only real question is whether the size should decrease from top to bottom and from left to right, or the reverse; and whether more recent dates should be listed at the top or at the bottom and at the left or at the right. The general rule is that the magnitude shall increase from left to right or from top to bottom, and that time series shall be arranged with the more recent items at the right or bottom. In some instances, particularly in the current publication of certain time series of business and economic statistics, the reverse order is

followed; but the practice has little to commend it, except the placing of current items in the most prominent position. For simple frequency series which are to be subjected to charting or computation, there is some advantage in having the variable increase toward the top because this is the positive direction on a chart.

For categorical series the case is not so clear. Sometimes the arrangement is made according to the size of the item, sometimes according to generally accepted notions of the importance of the objects listed, sometimes according to some external consideration pertaining to the order of arrangement such as the geographical location of the states of the United States, sometimes alphabetically according to the names of the objects, and occasionally according to other schemes of minor significance.

Labels.—The title of the table requires special care in its preparation. It should state in full the essential facts concerning the content of the table—the facts common to all the items of the table—and usually should list these facts in the order of their importance. On the other hand, it should not be unnecessarily wordy; it must be sufficiently compact so that reference to it will be convenient. Included in a subordinate position in the title, or attached to it in footnotes, should be clear statements as to the source of the data—whether primary or secondary—and the units in which the items of the table are expressed. By exception, in case sources or units are different for different portions of the table, the appropriate statements should be included in the proper stubs and captions or attached to them in footnotes.

Moreover, if particular items of the table are exceptional in that they are derived from exceptional sources or pertain to objects not properly belonging to the general group under observation, each such item should have a footnote attached to it to give in full the special information necessary to a proper understanding of its irregular character. For example, if in a monthly record of prices of a particular commodity all but three months are covered by a single uniform source and those three items are obtained by interpolation from some other source or are estimated by an arithmetical formula, a footnote reference to those items should state this fact clearly.

The remarks which apply to the general title of the table hold for the stubs and captions, except that here the need of brevity is more urgent. It is frequently desirable to include in the stubs serial numbers for the rows from top to bottom, to facilitate

reference from the text; and, in a wide table, repetition of these numbers on the right edge facilitates tracing rows from left to right. Likewise, columns should often be designated by serial numbers or, if they are not too numerous, by letters. In giving the limits of the class intervals of a frequency series in the stubs (or captions) the exact interval should be unambiguous. For instance, in Table 23 it is uncertain whether the stubs state the lower limits or the upper limits or the midpoints of the several intervals, whereas Table 24 defines the intervals more precisely.

TABLE 23

DISTRIBUTION OF EMPLOYEES (PUDDLERS, LEVEL HANDED) IN PUDDLING MILLS IN 1931, ACCORDING TO EARNINGS*

Earnings[a]	Number of employees	Earnings[a]	Number of employees
55	28	75	39
60	21	80	7
65	29	85	1
70	86	90	3

* Derived from data in "Wages and Hours of Labor in the Iron and Steel Industry, 1931," Washington, U. S. Bureau of Labor Statistics, Bulletin 567, December, 1932, p. 83.
[a] Unit: cents per hour.

TABLE 24

DISTRIBUTION OF EMPLOYEES (ROLL ENGINEERS) IN BLOOMING MILLS IN 1931, ACCORDING TO EARNINGS*

Earnings[a]	Number of employees	Earnings[a]	Number of employees
70–75	6	95–100	8
75–80	10	100–110	3
80–85	1	110–120	4
85–90	9	120–130	5
90–95	4	130–140	1

* Source: "Wages and Hours of Labor in the Iron and Steel Industry, 1931," Washington, U. S. Bureau of Labor Statistics, Bulletin 567, December, 1932, p. 98.
[a] Unit: cents per hour. Lower limit inclusive.

Rulings.—Except for tables of the first order, a logical set of rulings is quite indispensable to efficiency in the use of tabulated material; and even for tables of the first order simple rulings are distinctly helpful, although discriminating use of spacing may render some ruled lines unnecessary. The rulings must be designed with a view to their serving as systematic guides to the reader as he enters the table in search of an item. Differences in

width, heaviness of face, or color of the lines should assist in distinguishing those divisions of the table between primary groupings from those between subordinate.

The rulings and spacings must guide the reader unerringly from the title of the table to the principal caption and principal stub, and then (in a complicated table of high order) to the subordinate caption and subordinate stub in which he is interested, and then to the item at the intersection of the particular column and row designated by the selected caption and stub. If this notion that the rulings and spacings are guides is kept in mind, many illogical schemes will be avoided, as well as many schemes which may be logical but are of little practical utility. With few exceptions, a double or heavy ruling should stand just beneath the title and bound the top of the table, and a ruling should set off the captions (or stubs) from the adjacent rows (or columns). A single line at the bottom of the table is usually sufficient. The extreme left and right edges of the table should be left open rather than being closed by vertical rulings.

Careful use of spacing assists the reader in using a table. Thus, if the table contains numerous rows between which there is no occasion to introduce rulings, the long columns may be broken by grouping the items through the widening of the space after every third, fifth, or tenth item. Similarly, if each item contains many digits, they should be grouped: 37884296 is less satisfactory than 37 884 296. Many additional detailed suggestions bearing upon the construction of tables will become apparent to the student as his experience with existing tabulations and the construction of his own tables develops. The essential principle is always the same: devices used in making a table should be directed to rendering the data readily accessible and to describing them unambiguously without sacrificing the compactness and simplicity of the table as a whole.

CHAPTER VI

SUMMARY TABLES

A statistical inquiry, unless its object is merely the collection and publication of primary data, ordinarily requires the organization of data into a more specialized form than a general table. Such specialized tabulations are essentially summary tables derived, or derivable, from general tables. The summary table ordinarily presents only a portion of the data of the basic general table, arranges those selected data in a form facilitating their comparison or study, or accompanies them by certain ratios or averages or other calculated statistical numbers. Thus the summary table is fundamentally an instrument of analysis: it is organized for the purpose of bringing out significant relations among the data or of placing emphasis upon particular items or groups of items. Unlike the general table it is not primarily a mere repository of statistical facts, although it customarily gives an acceptable record of those basic data which it reproduces from the general table.

A summary table may come into being in various ways. In a study based upon an original survey, the data collected may first be listed in a general table, and then selected portions of the data may be arranged and reproduced in summary tables. This practice is common in certain official publications, for example, the chief statistical publications of the United States Census. On the other hand, although this is a less satisfactory procedure from the point of view of the user of data, the items collected in an original survey may be entered at once—without first appearing in general tables—in summary tables. At best, this procedure is available only for cases in which the classification scheme is rather simple and the body of material to be presented is not large. And even in these cases the primary source may have *used* a general tabulation as an intermediate step in preparing the summary tables which alone are *presented.*

For most investigations in economic statistics, however, the investigator uses secondary data. His procedure normally consists in finding the desired data in an already published source or sources, preferably primary as noted above, and then compiling his

desired summary tables from the data given in the source. The data in the source may appear in general tables or, particularly if it is a secondary source, summary tables. The investigator will normally prefer to take his data from general tables in sources; because if he takes data from summary tables his selection of data is limited by whatever selection governed the organization of those summary tables. Moreover, the very form and content of such summary tables may give him a more or less serious bias in his own selection and organization of material, as would ordinarily not be the case if he quoted data only from general tables.

WORKING AND PUBLICATION TABLES

The form and arrangement of a summary table are controlled to some degree by the manner of its origin—whether it is derived from primary or from secondary data, and in the latter case whether the secondary data appeared in general or summary tables. But the use to be made of a summary table is a much more effective control. The two main classes of summary tables—according to their use—are *publication tables*, also called *presentation tables*, and *working tables*.

The working table is compiled as an intermediate step in a statistical analysis: it is an instrument in the hands of the investigator. The publication table is prepared for presentation along with, or including, the results of the analysis. It is indeed a part of the result, and it forms the tabular constituent of the report of an investigation. Like the textual part of the report, it is therefore designed to convey certain facts and to suggest or aid their interpretation. It is the end product, so far as tabulated material is concerned, of the statistical analysis; and it should therefore be prepared not only to ensure its accurate representation of facts but also with a view to its effect upon the mind of the reader. The working table, on the other hand, does not need to be planned with a view to its examination by a reader who may not be a trained statistician. It is merely a device to aid the statistician in his work, and its entire design is governed by considerations of the accuracy and economy of his operations. In planning it, he needs only to think of himself or his assistants and of the way in which it will fit into the program of analysis he has in view.

Requisites for summary tables.—The form of a working table may, therefore, differ widely from that of a publication table containing substantially the same data, but certain chief requirements of good tabulation practice apply to both types. To a large

extent, these are the same requirements applicable to the preparation of general tables. Summary tables of both types require: an adequate descriptive title, logical selection and arrangement of stubs and captions, convenient guides in the form of rulings and spacings, clear specification of units, complete reference to the source of every item, comprehensive and unambiguous footnotes on all exceptional items.

The statistician will understand readily that such requirements must be met for publication tables, but he will be tempted to ignore some or nearly all of them in preparing his working tables. Despite emphatic admonition in this text, he will probably have to learn by sad experience: his own annoyance and waste of effort in trying to use, in the final stages of preparing reports of his investigations, working tables in which he has failed to meet one or more of the above requirements will ultimately convince him that such neglect leads to a heavy penalty in time and work.

In addition to the requirements which pertain also to general tables, summary tables must ordinarily meet additional special requirements. If several presentation tables of like content and like import are to appear in a report, they should be organized along similar lines. A wise degree of standardization is desirable. Likewise, if the summary tables of a report are linked to charts shown in the report, much is gained by an adaptation of the table plans to the chart plans. These requirements apply also to working tables; for such tables, moreover, an adaptation of the plan of working tables used in earlier stages of a study to that of those used at later stages generally promotes economy and accuracy in analysis and interpretation. These points suggest that the analyst should exercise foresight in planning the entire scheme of tabulation—both for working and for presentation tables—of his analysis. This desideratum cannot always be realized, of course, and the statistician not infrequently must feel his way in an analysis and keep his plan of tabulation sufficiently flexible to meet phases of the analysis which he could not foresee.

PREPARATION OF WORKING TABLES

The plan of a working table is usually controlled by the specific manner in which the data will be used. By exception, the tables compiled in the initial stage of an inquiry may be set up with a view merely to ease and accuracy in transcribing the desired data from sources. Particularly if the source data appear in a somewhat complicated form, the transcription operation may be suffi-

ciently involved to force the adaptation of these first working tables to the transcription process. In such cases these first transcription sheets serve merely as the basis for compiling the next group of working tables, which can then be designed more definitely with a view to their use; and, even in the more difficult transcription operations, some attention can be given to arranging the transcription tables for further use.

Blank forms for transcription.—Some compromise, dictated by practical considerations in the particular statistical inquiry, is accordingly made between adapting the first working tables—the transcription sheets—to the form in which the data appear in the source or sources and to the use to be made of the data in the next stage of the analysis. The compromise effected aims at reducing the labor and risk of error incident to handling numerous sheets, in the transcription operation and in subsequent work. If several series of somewhat similar nature are to be transcribed for use in a single investigation, some uniformity is desirable in the forms used for the various series; like material should be arranged in like manner, to facilitate further use.

A satisfactory *transcription form* should have generous spaces, so that the individual items stand out clearly and so that revisions and corrections can be entered without undue crowding and without erasing the first entries. There should be ample space, clearly set apart, for adequate descriptive designation of each item or series, including units and source references, and for essential footnotes. All this merely provides room for meeting the requirements of good tabulation: the descriptive designation should state completely and specifically what the items are; the title or footnotes should contain a specification of the unit in which each item or series is expressed; the exact source of each item should be clearly given in the footnotes, and the reference to source should state the author, title, edition, publisher, place of publication, date of publication, volume, and page of a book; or the name, issue date, and page of a periodical. Where the several items are taken from different sources, or from different issues of the same source, the source reference should be detailed enough to locate definitely the source of each item. Pertinent footnotes or descriptive comment appearing in the source should be carried over in footnotes to the transcription form. The final transcribed sheet should give a sufficiently complete description of all the data thereon to render reference to the source unnecessary when the sheet is being used for analysis or interpretation. The form

should be designed to give room for entering all these facts; at the same time compactness of form, and sufficient uniformity of forms which are to include like data, should be sought so far as compatible with the absolutely essential requirement that there be no crowding. (For further comment on transcription forms, see Appendix B.)

Purpose of a working table.—As noted above, the plan of a working table, even a transcription sheet, is usually governed by its use. For example, if a working table is designed merely to assemble data for charting, the chief considerations are accuracy, simplicity, and accessibility. Such a table should contain only the data to be charted, and not a mass of other material; and it should be arranged to facilitate prompt and accurate transfer of its items to the chart. Evidently, foresight as to the nature and the arrangement of the chart is helpful in designing such a table. Tables 2, 3, and 5 might be regarded as simple working tables of this sort, made for the sole purpose of plotting Charts 1 to 4. Table 25 is a somewhat less simple case, in which the data of the source have been specially classified in the working table, with a view to preparing Chart 9 (page 97).

A working table which is a step in a computation operation is ordinarily less easily designed. The design of such a table aims to promote speed and accuracy in computation. Usually a table of this sort carries spaces not only for entering the basic data but also for entering the results subsequently computed. Tables 26 to 28 are of this sort. Table 26 is an example of a computing table in which the basic data do not appear: a working table at a previous stage of the analysis has derived, from the basic data or from the results of a still earlier working table, the "median link relatives," and those become the "basic data" of Table 26. In Table 27 the basic data appear in the column adjacent to the stubs; all other columns, and the bottom cell of the table, give results of computations. In Table 28 the basic data appear in all but the bottom row; if the totals (third and seventh rows) had not been transcribed as basic data, these also would be computed figures.

It is desirable, if not essential, to anticipate the entire process of computation in planning the blank form for such a working table. Otherwise space may not be available for some of the computed figures, or such spaces may not be conveniently located with respect to the items from which computations are to be made, or the basic items may not be arranged in the order which best fits the work of computation. The student will learn, from experience

with clumsy arrangements of computation sheets, how to avoid the major obstacles to economical and accurate work. Such a table

TABLE 25

CLASSIFICATION OF STATES ACCORDING TO NUMBER OF MOTOR CARS REGISTERED IN 1936*

State	Under 0.1	0.1 to 0.2	0.2 to 0.4	0.4 to 0.6	0.6 to 0.8	0.8 to 1.0	Over 1.2[a]
Alabama			x				
Arizona		x					
Arkansas			x				
California							x
Colorado			x				
Connecticut			x				
Delaware	x						
Dist. of Columbia		x					
Florida			x				
Georgia				x			
Idaho		x					
Illinois							x
Indiana						x	
Iowa					x		
Kansas				x			
Kentucky			x				
Louisiana			x				
Maine		x					
Maryland			x				
Massachusetts						x	
Michigan							x
Minnesota					x		
Mississippi			x				
Missouri						x	
Montana		x					
Nebraska				x			
Nevada	x						
New Hampshire		x					
New Jersey						x	
New Mexico		x					
New York							x
North Carolina				x			
North Dakota		x					
Ohio							x
Oklahoma				x			
Oregon			x				
Pennsylvania							x
Rhode Island		x					
South Carolina			x				
South Dakota		x					
Tennessee			x				
Texas							x
Utah		x					
Vermont	x						
Virginia				x			
Washington				x			
West Virginia			x				
Wisconsin						x	
Wyoming	x						

* Unit: million cars. Source: "Automobile Facts and Figures," 1937 edition, Automobile Manufacturers Association, New York, p. 17.

[a] The column of items 1.0 to 1.2 has been omitted since there are no data for this class interval.

needs to be as compact as possible without scrimping space needed for labels and figures, and to present the results of computation

clearly and in an order facilitating use in discussion or in passing to the next stage of working tables. Simplicity and orderliness in arrangement of the data and of the successive computation steps will promote speed and accuracy in calculation.

Obviously, the entire *computation form* should ordinarily be laid out and labeled in detail before any figures—even the basic figures—are entered. Effective planning will render such a skeleton

TABLE 26

COMPUTATION OF THE ADJUSTED INDEXES OF SEASONAL VARIATION FOR THE VALUE OF CONTRACTS AWARDED IN THE UNITED STATES (F. W. DODGE CORPORATION), BASED UPON LINK RELATIVES FOR THE INTERVAL MARCH, 1919, TO FEBRUARY, 1931*

Month	Median link relative	Logarithm	Corrected logarithm	Cumulative corrected logarithm	Anti-logarithm	Index
January	0 920	9 9638	9 9648	9 9648	92 2	74
February	1 105	0 0433	0 0443	0 0091	102 1	83
March	1 305	0 1156	0 1166	0 1257	133 6	108
April	1 150	0 0607	0 0617	0 1874	154 0	124
May	0 965	9 9845	9 9855	0 1729	148 9	120
June	0 955	9 9800	9 9810	0 1539	142 5	115
July	0 940	9.9731	9 9741	0 1280	134 3	109
August	0 945	9 9754	9 9764	0 1044	127 2	103
September	0 995	9.9978	9 9988	0 1032	126 8	102
October	0 915	9 9614	9 9624	0 0656	116 3	94
November	0 920	9 9638	9 9648	0 0304	107 3	87
December	0 930	9 9684	9.9694	9 9998	100 0	81
Total		9 9878			1485.2	
		−0 0122				
Average		−0 0010			123 8	

* Data are based upon averages per working day as described in the *Review of Economic Statistics*, Vol XIII, 1931, pp. 68–75.

layout possible in most cases, although the statistician may find, particularly in undertaking an unfamiliar type of analysis, that he needs to experiment by actually entering data in tentative computation sheets until he sees how his space requirements work out. After such experimentation, however, he should settle down to a well-designed set of computation forms and use them for the entire job. Even in the very simplest schemes of computation, any work sheets which are slovenly or poorly planned should be avoided. In connection with less simple schemes, even those only moderately elaborate, the forms should ordinarily be planned with great care; and the copies used should be printed, or otherwise

produced with mechanical uniformity, upon a stout paper which will take ink clearly.

Beyond the foregoing basic rules, not much can be suggested in general about the construction of working tables for computation. Each type of analysis presents its own peculiar difficulties and requires its own plan. The statistician will acquire skill as he acquires experience; he should always aim at simplicity and

TABLE 27

COMPUTATION OF THE ARITHMETIC AVERAGE WAGES OF MALES UNDER 16 YEARS OF AGE IN ALL OCCUPATIONS IN FOUNDRIES AND METALWORKING INDUSTRIES IN CENTRAL STATES IN 1900*

Dollars per week	Number (f)	x	xf	
			−	+
2 50–2 99	13	−4	52	
3 00–3 49	58	−3	174	
3 50–3 99	102	−2	204	
4 00–4 49	40	−1	40	
4 50–4 99	55	0		
5 00–5 49	46	1		46
5 50–5 99	8	2		16
6 00–6 49	32	3		96
6.50–6 99	32	4		128
7.00–7 49	10	5		50
7 50–7 99	11	6		66
8.00–8 49	11	7		77
8 50–8 99	8	8		64
Total	426		−470	+543

Mean = $4\ 75 + {}^{73}/_{426} \cdot 50 = 4\ 84$

* Source "Twelfth Census of the United States, 1900, Special Report on Employees and Wages," Washington, U. S. Census Office, 1903, p. 669.

economy without seeking them to the point of endangering accuracy. (For further comment on computation tables, see Appendix B.)

PREPARATION OF PUBLICATION TABLES

Many of the fundamental rules applicable to the preparation of working tables apply also in preparing publication tables. Differences in detail arise, however, chiefly because a publication table is addressed to individuals not familiar with, and perhaps not interested in, the basic data or the statistical methods of analysis. The working table, on the other hand, is a device of analysis for the

statistician who is acquainted with both the data and the method of the analysis.

Furthermore, the intended use of the publication table as a basis for textual discussion or interpretation in the report of the analysis, rather than as a step in computation or other analysis, governs to some degree its plan. The very finality of a publication table implies certain features usually not found in a working table. In this connection, a particular class of working tables should be

TABLE 28

COMPUTATION OF THE RATIO OF TOTAL RESERVES TO DEPOSIT AND FEDERAL RESERVE NOTE LIABILITIES COMBINED, FOR THE COMBINED FEDERAL RESERVE BANKS, AT TWO DATES IN 1936*

Item	September 30	December 31
Liability on		
Deposits	6,843,512	7,108,919
Federal reserve notes in circulation	4,049,143	4,283,537
Deposits and notes combined	10,892,655	11,392,456
Reserves		
Gold certificates on hand and due from United States Treasury	8,384,683	8,851,880
Redemption fund—Federal reserve notes	12,428	12,741
Other cash	261,445	256,534
Total reserves	8,658,556	9,121,155
Ratio of total reserves to deposit and Federal reserve note liabilities combined	79.5	80.1

* Unit for deposits, notes, and reserves, thousand dollars; for ratio, 1 per cent. Source: "Twenty-Third Annual Report of the Board of Governors of the Federal Reserve System," Washington, 1937, pp. 74, 75.

mentioned: the working tables developed in or near the final stage of an analysis are likely to be very similar in all important details to the publication tables used in the report. The statistician himself finds it advantageous to arrange the summary tables of the last stages of his analysis upon about the same plan which he will subsequently use in publication. We may fairly say that, as he passes from stage to stage of his analysis, the publication point of view is more and more in his mind; and his working tables correspondingly approach more and more closely to the form he is finally to use in publication.

A summary table for presentation usually contains only a limited amount of material, specifically chosen to serve as the factual basis of particular passages in the accompanying text, or selected and arranged to convey by itself a definite principle or

conclusion. The arrangement of the material within the table is dictated partly by considerations of compactness and accessibility but mainly by a knowledge of the relations and comparisons which the table is intended to bring out. In outlining a table to be presented for a special purpose, the distribution of emphasis can be regulated in part by proper utilization of those positions in the

TABLE 29

ANALYSIS OF OPERATING REVENUES, ALL DISTRICTS, CLASS I STEAM RAILWAYS OF THE UNITED STATES, FOR THE YEAR ENDING DECEMBER 31, 1936 (EXCLUDING SWITCHING AND TERMINAL COMPANIES)*

Account	Amount	Average per mile of road	Per cent of total revenues
Operating revenue			
Freight	$3,302,894,429	$13,943	81.50
Passenger service train revenue	591,212,150	2,496	14 59
Switching	57,563,881	243	1.42
Water transfers—freight	576,318	2	0 01
Water transfers—passenger	1,312,230	6	0 03
Water transfers—vehicles and livestock	3,062,824	13	0 08
Water transfers—other	889,078	4	0 02
Total rail-line transportation revenue	3,957,510,910	16,707	97.65
Freight	5,646,267	24	0 14
Passenger	235,049	1	0 01
Excess baggage	275	[a]	[b]
Other passenger service	21,349	[a]	[b]
Other	161,932	1	[b]
Total water-line transportation revenue	6,064,872	26	0 15
Total incidental operating revenue	80,655,937	340	1.99
Total joint facility operating revenue	8,502,420	36	0 21
Total railway operating revenue	4,052,734,139	17,109	100.00
Mileage operated, miles	236,878[c]		

* Source: "Statistics of Railways in the United States, 1936," Washington, Interstate Commerce Commission, 1937, p S-71, with omission of detail rows.

[a] Less than $1. [b] Less than 0.01 per cent. [c] Represents average mileage of road operated during year.

(Attention is called to the existence in the stubs of a main classification, a primary subclassification, and a secondary subclassification, to the emphasis upon the primary subclassification afforded by the double horizontal rulings, to the manner in which the units—$ and miles—are specified, and to the computed comparison items of the "Average per mile of road" and "Per cent of total revenues" columns.)

table which have greater prominence. Moreover, desired comparisons between particular items or series can be facilitated by placing such items or series as nearly as possible in adjacent rows or columns. In many instances the purpose of the table is furthered not only by an appropriate arrangement of the given items but by

the inclusion of certain derived items, such as per cent ratios or averages.

Illustrations of summary tables.—Tables 29 to 36 illustrate cases of publication tables and bring out many practical points—some of which are mentioned in the commentaries below the tables—concerning tabulation procedure. The student will profit by examining these and other summary tables in this text and in other publications and by familiarizing himself with the devices used to meet the needs of particular cases. This discovery of the purpose and manner of organization of such a table is not sufficient; the student should actively undertake to recast and improve tables which he encounters in his reading. By such a practice he will train himself more effectively to use wise tabulation plans in his own work.

TABLE 30

VALUE OF MERCHANDISE TRADE OF THE UNITED STATES, BY GROUPS OF COMMODITIES*

Year[c]	Imports[a]					Exports[b]				
	Group A[d]	Group B	Group C	Group D	Group E	Group A	Group B	Group C	Group D	Group E
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1926 (total)	1,792	540	418	804	877	1,261	335	503	656	1,957
1927 "	1,601	505	451	750	879	1,193	421	463	700	1,982
1928 "	1,467	550	406	763	906	1,293	295	466	716	2,260
1929 "	1,559	539	424	885	994	1,142	270	484	729	2,532
1930 "	1,002	400	293	608	757	829	179	363	513	1,898
1931 "	642	305	222	372	549	567	127	247	318	1,120
1932 "	358	233	174	217	341	514	89	152	197	624
1933 "	418	216	201	292	322	591	48	155	237	617
1934[e] "	461	254	264	307	350	653	59	168	342	879
1935 "	582	322	319	410	406	683	59	157	350	994

* Unit: one million dollars. Source: "Statistical Abstract of the United States, 1936" Washington, U S Department of Commerce, 1936, pp. 438–439.

[a] Import data are "general imports" through 1933, "imports for consumption" thereafter

[b] Export data are United States merchandise.

[c] Data are for calendar years.

[d] Group A: crude materials, Group B: crude foodstuffs; Group C: manufactured foodstuffs, including beverages, Group D. semimanufactures, Group E: finished manufactures.

(Attention is called to the use of letters as captions, with those letters defined in footnote [d]. This practice does away with the necessity of cramping long captions into the narrow columns. The individual columns are not distinguished by the letters but by the numerals which appear in parentheses. It should be observed also that the main captions—"Imports" and "Exports"—are not boxed in by rulings; the double vertical ruling is relied upon to set off the columns of the first main caption from those of the second.)

TABLE 31

DEMAND DEPOSITS OF NATIONAL BANKS, JUNE 30, 1936*

	Deposits of individuals, partnerships, and corporations (A)	United States Government deposits (B)	State, county, and municipal deposits (C)	Deposits of other banks in the United States (except private banks and American branches of foreign banks) (D)	Deposits of private banks and American branches of foreign banks (E)	Deposits of banks in foreign countries (including balances of foreign branches of other American banks, but excluding amounts due to own foreign branches) (F)	Certified and cashiers' checks (including dividend checks), letters of credit and travelers' checks sold for cash, and amounts due to federal reserve bank (transit account) (G)	Total
Central reserve cities								
New York	2,538,765	81,651	194,392	875,576	55,201	164,227	129,988	4,039,800
Chicago	1,133,618	92,379	208,216	548,575	1,441	4,659	19,526	2,008,414
Total central reserve cities	3,672,383	174,030	402,608	1,424,151	56,642	168,886	149,514	6,048,214
Total other reserve cities	4,530,232	365,830	578,744	2,026,830	13,405	24,290	116,691	7,656,022
Total all reserve cities	8,202,615	539,860	981,352	3,450,981	70,047	193,176	266,205	13,704,236
Total country banks	3,463,257	152,667	805,132	321,733	1,572	1,008	87,439	4,832,808
Total United States	11,665,872	692,527	1,786,484	3,772,714	71,619	194,184	353,644	18,537,044

* Unit: one thousand dollars. Source: "Seventy-Fourth Annual Report of the Comptroller of the Currency (covering the year ended October 31, 1936)," Washington, U. S. Treasury, 1937, pp. 468–471, with omission of detail rows.

(Attention is called to the unusual length of the captions which are necessary to describe the kinds of deposits. The use of letters to designate the different detail columns should also be noted; this practice facilitates the reading of a table which extends over several pages, as does the above table when its many rows are retained. In this table, the plan, formerly standard and still very common, of entering totals at the right and bottom was followed.)

TABLE 32

SERIES INDICATIVE OF VEHICLE AND EQUIPMENT PRODUCTION*

Date	Automobiles produced[a]		Automobile tires[b]				Domestic railroad car orders[d]		
			Production		Stocks[c]				
	Passenger cars	Trucks	Pneumatic casings	Inner tubes	Pneumatic casings	Inner tubes	Locomotives	Freight cars	Passenger cars
	Thousand cars	Thousand trucks	Thousand casings	Thousand tubes	Thousand casings	Thousand tubes	One locomotive	One car	One car
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1926 (monthly average)	315.33	43 08	5,010	6,192	10,599	16,938	108	5,586	156
1927 (monthly average)	244 71	38 73	5,296	5,905	10,832	15,873	61	6,001	134
1928 (monthly average)	317 95	45.28	6,282	6,743	11,343	15,530	50	4,267	161
1929 (monthly average)	382 28	64 25	5,814	6,170	14,233	16,016	101	9,268	192
1930 (monthly average)	232 06	47 60	4,299	4,594	11,599	12,572	37	3,863	56
1931 (monthly average)	164 42	34.72	4,095	4,097	9,265	9,505	15	907	1
1932 (monthly average)	94 62	19.60	3,357	3,114	7,715	7,339	1	164	3
1933 (monthly average)	131 13	28 88	3,781	3,575	7,525	6,884	4	140	0
1934 (monthly average)	181 49	47 93	3,936	3,852	9,903	8,860	15	2,051	32
1935 (monthly average)	271 02	57 89	4,030	3,990	9,681	9,023	7	1,558	5

* Source: columns 1–2, *Survey of Current Business*, 1936 Supplement, p. 147; columns 3–6, *ibid.*, p. 134; columns 7–9, *ibid.*, p. 149.

[a] Automobile production data (for the United States) represent complete manufacturers' sales, commonly referred to as *production*, as compiled since 1921 by the U. S. Department of Commerce, Bureau of the Census, in cooperation with the Automobile Manufacturers Association. For further details regarding the figures, see *Survey of Current Business*, 1936 Supplement, p. 183.

[b] Computed by U. S. Department of Commerce, Bureau of Foreign and Domestic Commerce. For additional details see *Survey*, 1936 Supplement, p. 180.

[c] Based on end-of-month figures. [d] Comprises new orders placed by all buyers. See *Survey*, 1936 Supplement, p. 184.

(Attention is directed to the arrangement of the main and subordinate captions, to the specification of the units at the heads of the several columns, to the fact that certain footnotes are attached to the main captions and another to a subordinate caption, to the specification of the source by use of the column numerals, and to the absence of double rulings within the table.)

TABLE 33

EMPLOYMENT AT IRON-ORE MINES AND BENEFICIATING PLANTS IN THE UNITED STATES, QUANTITY AND TENOR OF ORE PRODUCED, AND AVERAGE OUTPUT PER MAN, 1926–1935*

Year	Employment					Production										
	Average number of men employed	Time employed				Crude ore (partly estimated), gross tons	Merchantable ore			Average per man (gross tons)						
		Average number of days	Total man shifts	Man-hours			Gross tons	Iron (natural) contained		Crude ore (partly estimated)		Merchantable ore				
				Average per day	Total			Gross tons	Per cent	Per shift	Per hour	Per shift	Per hour	Iron (natural) contained	
														Per shift	Per hour
1926	34,399	273	9,395,178	9.0	84,225,524	75,943,775	67,623,000	34,099,262	50 43	8.083	0 902	7.198	0.803	3.629	0.405
1927	34,755	264	9,177,979	8.9	82,004,761	69,923,057	61,741,100	30,879,989	50 02	7 619	0.853	6.727	0.753	3.365	0.377
1928	30,238	265	8,008,647	8 9	71,403,631	70,940,916	62,197,088	31,149,584	50.08	8.858	0.994	7.766	0.871	3.889	0.436
1929	30,763	281	8,638,234	8 9	77,111,086	83,164,881	73,027,720	36,637,660	50 17	9.628	1.079	8.454	0.947	4.241	0.475
1930	30,975	259	8,037,096	8.9	71,620,115	68,551,913	58,408,664	29,212,457	50.01	8.529	0.957	7.267	0.816	3.635	0.408
1931	22,867	201	4,596,504	8.9	40,928,283	35,563,994	31,131,502	15,625,050	50.19	7.737	0.869	6.773	0.761	3.399	0.382
1932	12,649	145	1,828,002	9 0	16,427,009	11,181,678	9,846,916	4,948,243	50 25	6.117	0.681	5.387	0.599	2.707	0.301
1933	15,125	140	2,121,494	8.5	17,931,479	21,225,958	17,553,188	8,777,574	50 01	10.005	1.184	8.274	0.979	4.137	0.490
1934	16,513	193	3,186,232	8.0	25,478,440	28,252,926	24,587,616	12,384,257	50.37	8 867	1.109	7.717	0.965	3.887	0.486
1935	14,987	219	3,277,995	8.0	26,281,693	35,367,816	30,540,252	15,361,718	50.30	10.789	1.346	9.317	1.162	4.686	0.585

* Exclusive of ore containing 5 per cent or more manganese. Source: "Minerals Yearbook, 1937," Washington, U. S. Department of the Interior, 1937, p. 600.

(This table illustrates the handling of a fairly complicated subclassification scheme in the captions. The rulings, although simple in plan, effectively supplement the captions in organizing the data of the table.)

TABLE 34

COKE PRODUCED IN THE PRINCIPAL COUNTRIES OF THE WORLD, 1933–1936*

Country	1933	1934	1935	1936
Australia				
New South Wales	481,026	699,673	871,644	[a]
Queensland	15,337	26,067	25,276	[a]
Belgium	4,694,130	4,601,950	4,678,400	5,050,000
Bulgaria	628	935	1,705	1,683
Canada	1,228,246	1,658,691	1,663,515	1,809,204
China (exports)	1,709	6,531	7,246	11,422
Chosen	220,500	246,900	[b]	[a]
Czechoslovakia	1,259,381	1,344,786	1,553,869	1,955,000
France	6,787,600	7,293,110	7,078,000	7,030,000
Germany	21,153,744	24,484,890	29,556,269	35,861,000
Saar	1,880,000	2,180,000	[c]	[c]
Great Britain[d]	8,919,540	11,697,111	12,131,081	[a]
Hungary	9,163	19,086	22,981	30,704
India, British[e]	1,247,451	1,541,487	1,795,178	[a]
Indo-China	360	285	260	[a]
Italy	729,966	817,243	998,382	[a]
Japan				
Manufactured coke	[b]	[b]	[b]	[a]
Natural coke	370,785	367,236	396,214	[a]
Mexico	251,604	275,176	489,047	[a]
Netherlands	2,609,373	2,779,378	2,878,191	[a]
Peru	[b]	[b]	[b]	[a]
Poland	1,170,717	1,333,493	1,386,716	1,615,598
Rhodesia, Southern	31,798	55,979	39,239	[a]
Rumania	7,150	31,914	45,920	63,391
Spain	427,453	485,634	[b]	[a]
Straits Settlements	7,860	8,549	[b]	[a]
Sweden	103,336	107,370	114,464	115,430
U. S. S. R. (Russia)	10,225,000	14,221,000	16,752,000	19,883,000
Union of South Africa	75,456	72,969	64,782	75,459
United States	25,028,365	28,867,897	31,879,449	41,979,921
Total	90,218,000	106,505,000	115,969,000	[a]

* Unit: metric tons. Gashouse coke is not included. Source: "Minerals Yearbook, 1937," Washington, U S Department of the Interior, 1937, p. 926.

[a] Data not available

[b] Estimate included in total.

[c] Beginning with March, 1935, production of the Saar is included with that of Germany.

[d] In Great Britain the production of gashouse coke (including breeze), not included above, is especially important and was as follows: 1933, 11,657,081 tons; 1934, 12,038,825 tons; 1935, 12,181,117 tons.

[e] Data shown represent total "hard" coke manufactured. In addition, the following quantities of "soft" coke were made at collieries: 1933, 837,393 tons; 1934, 874,901 tons; 1935, 904,840 tons.

(The extensive use of footnotes to indicate breaks in homogeneity is the most striking feature of this tabulation.)

TABLE 35

FREIGHT CARS OWNED AND ON ORDER, CLASS I RAILROADS OF THE UNITED STATES, JANUARY–DECEMBER, 1937*

1937	Owned		Unfilled orders		
	Number (thousand cars)	Capacity (million pounds)	Total	Equipment manufacturers	In railroad shops
January	1,741	170,109	33,608	27,414	6,194
February	1,738	169,887	39,729	31,214	8,515
March	1,733	169,682	44,708	34,314	10,394
April	1,732	169,665	46,197	35,814	10,383
May	1,731	169,839	44,397	31,802	12,595
June	1,729	169,883	41,895	29,577	12,318
July	1,730	170,102	37,411	23,952	13,459
August	1,732	170,409	31,123	19,525	11,598
September	1,732	170,585	24,225	14,155	10,070
October	1,732	170,791	18,231	9,725	8,506
November	1,735	171,085	12,511	5,463	7,048
December	1,731	170,809	7,904	2,896	5,008

* Data are for end of month. Source: *Survey of Current Bussness*, March, 1938, p 95.

(The fact that the main captions are "Owned" and "Unfilled orders" stamps this as essentially a general table, for the above arrangement is not designed to facilitate comparison.)

TABLE 36

METHOD OF COMPUTING VELOCITY OF BANK DEPOSITS*

Week ended	Debits to individual accounts (1)	Time deposits ÷ 26 (2)	Government withdrawals (3)	Revised debits (4)	Total debits each month (5)	Number of working days in each month (6)	Average daily debits (7)	Annual rate of debits (8)	Net demand deposits (9)	Net due to banks (10)	Revised demand deposits (average for month) (11)	Annual rate of turnover of deposits (12)
1922												
January 4	$4,529,355	$7,120	...	$4,522,235	.....	..	.....	.....	$3,866,822	$804,960		
11	4,592,367	7,370	$ 5,884	4,579,113	.....	..	.....	.....	3,850,902	799,187		
18	4,766,247	7,196	16,884	4,742,167	.....	..	.....	.....	3,788,338	781,546		
25	3,933,296	7,367	76,233	3,919,696	$18,571,486	25	$742,859	$224,343,418	3,754,903	782,753	$3,023,130	74.2
February 1	4,233,272	7,333	..	4,225,939								

* Case for 42 New York City Reporting Banks. 000 omitted except in columns 6 and 12. Source: *Journal of the American Statistical Association*, 1923, p. 730.

Time deposits turn over about twice a year, and checks drawn against time deposits each week therefore amount to about 1/26 of the amount of time deposits.

Items of columns 2 and 3 are subtracted from items of column 1 to yield items of column 4.

The monthly figure of column 5 is obtained by including debits of the weeks at the beginning and the end of the month in proportion to the number of working days of such weeks as fall in the month.

Column 8 is obtained by multiplying the items of column 7 by 302, the number of working days in the year.

Column 10 shows the excess of *due to banks* over *due from banks*, and is to be subtracted from column 9 to yield column 11.

Column 12 is the ratio of column 8 to column 11.

CHAPTER VII

CHARTING: CATEGORICAL SERIES

THREE METHODS OF SUMMARIZING DATA

Little benefit can be derived from examination of a mass of statistical material unless it is summarized and digested. Three chief methods exist for analyzing statistical material with a view to extracting from the entire body of data those facts and relations which are essential for a particular purpose. The first of these methods consists in constructing one or more *summary tables;* and, as appears from the discussion and illustrations of the previous chapter, the effectiveness of this process is narrowly limited. A second method is *arithmetic analysis*, which comprises the more or less lengthy and complicated computations by which are derived the numerical characteristics (summary descriptive numbers) of a particular series or the measurements of relation between several series. The methods available for arithmetic analysis are diverse and powerful, and numerous such methods, together with the difficulties and dangers attending their use, are discussed in subsequent chapters. The third method is *charting*, which is treated in the present chapter and the two which immediately follow.

The statistical chart is in some respects the pictorial equivalent of the summary table. It aims to present, graphically, certain portions of the material in such manner that the eye can quickly and accurately assess their meanings and their relations to each other; and, with rare exceptions, it conceals or ignores portions of the numerical detail or sacrifices some of the precision actually present in the data. One advantage of the table over the chart is that the data can be read more promptly and precisely from the table than from the chart. The great advantages of the chart, as compared with the corresponding table, are its capacity for the simultaneous presentation and comparison of a considerable group of facts and its flexibility in respect to the selection of data and apportionment of emphasis.

The obvious purpose of a chart is to render visible at a single glance, or upon very brief examination, the essential facts concern-

ing a single series, or the important relations between several series. This purpose clearly cannot be accomplished by a table, because the mind of the ordinary reader cannot secure an accurate and comprehensive view of even a small body of data. Whereas tabular presentation successfully indicates the major properties and the more evident relations among the statistical items of a group, it cannot give an adequate notion of any considerable number of such properties and relations nor can it promptly present numerous important but minor features of the material. Such properties and relations as cannot be emphasized in the table can often be displayed advantageously by properly constructed charts.

The chart, like the working table, serves as an aid to the statistician in studying his data. As in the case of summary tables, there are working charts and publication charts. The student should form the habit of making liberal use of charts in his work of analysis, and should expect to construct and examine many working charts which he does not contemplate finishing into publication charts. Particularly when he undertakes the arithmetical analysis of statistical series, he should almost invariably have before him a graphic record of his data in deciding upon methods of computation. By following such a practice he will often protect himself against carrying out elaborate calculations which would lead to results having little or no significance.

As respects the technique of chart making, little need be said here. Ordinarily working charts are made in pencil, and the lettering is not so carefully done as that of publication charts. Neatness and orderliness are, however, essential for working charts as well as for the more formal publication charts. (See Appendix B for brief suggestions on charting.)

LIMITATIONS OF GRAPHIC PRESENTATION

A first consideration controlling the kind of chart selected for a particular series is the classification of that series in one of the three groups: categorical, time, and frequency. Certain general limitations, however, apply to charts of all sorts. A chart appeals to the eye, and its meaning is determined visually. The design and construction of a chart must therefore take account of the limitations upon the sense of vision. Professor Bowley states that the eye can most accurately judge *distances* or differences in distance, and that the visual estimate of distances is aided, often without the conscious effort of the reader, by somewhat less

accurate but moderately reliable observations of *ratios* and *angles*.[1] These are the principal facts which the eye can quickly note concerning lines, points, and their relations to each other; and the chief additional devices by which the chart maker can convey information are differences in color or shading. Difficulties in reproduction obstruct the use of color in charts intended for presentation, and the labor of preparing different degrees of shading is an obstacle to the use of this device in the working charts used in the course of a study. The eye can accurately appraise only very few features of a diagram, and consequently a complicated or confusing diagram will lead the reader astray. The fundamental rule for all charting is to use a plan which is simple and which takes account, in its arrangement of the facts to be presented, of the above-mentioned capacities of the eye.

KINDS OF CHARTS FOR CATEGORICAL DATA

Charting is the method of analysis especially appropriate for categorical series. The individual facts and the comparisons in such series are frequently such that tabular summarization is not effective. Variations among the items are ordinarily so extensive that arithmetic summarization leads to results which are difficult to interpret, if not actually misleading. Hence graphic methods have a highly important place in the presentation and discussion of categorical data.

The *bar diagram* is the simplest and most adaptable general-purpose chart. Although any series, almost without exception, can be represented by a bar diagram, this type of chart is peculiarly satisfactory for categorical series. Charts 5 and 6 are examples of bar diagrams, one with vertical bars and the other with horizontal bars. No general rule can be given for preferring the horizontal or vertical position. To most readers the vertical bars give a more ready and more accurate impression of comparative size, but the horizontal bars lend themselves more easily to labeling with descriptive titles or numerical data.

Whichever plan is used, all bars should start from a single base line, a convenient scale should be indicated, and each bar should be labeled clearly. Sometimes the labels are placed upon the individual bars, as in Chart 7, but this arrangement is troublesome whenever some of the bars are short. More frequently they appear at the zero end of the bars, just outside the base line; and,

[1] "Elements of Statistics," p. 131.

in that position, they should not be so prominent because of heavy lettering as to give false impressions concerning the actual length of the bars (see Chart 6.). They should not be attached to the other extremities of bars (Chart 8), lest they lead to confusing illusions concerning the length of the bars. The order of arrange-

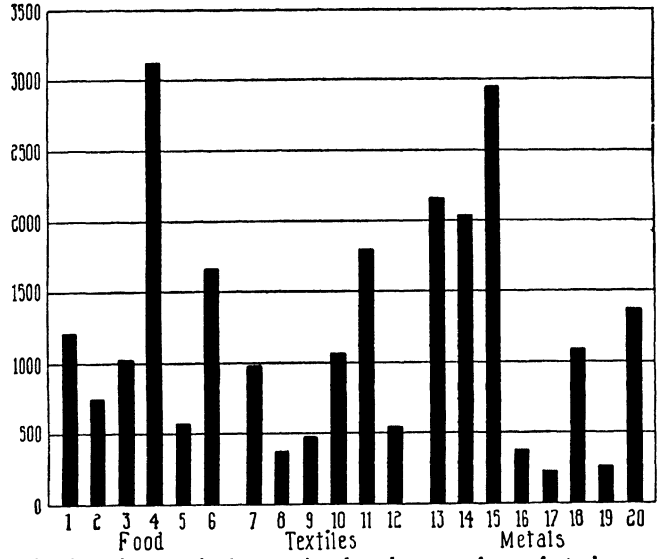


CHART 5.—Gross income of subgroups in selected groups of manufacturing corporations in 1934.

1. Bakery and confectionery products
2. Canned products
3. Mill products
4. Packing-house products
5. Sugar
6. Other food products
7. Cotton goods
8. Woolen and worsted goods
9. Silk and rayon goods
10. Other textiles
11. Clothing
12. Knit goods
13. Iron and steel
14. Machinery
15. Motor vehicles
16. Household equipment
17. Office equipment
18. Building material and hardware
19. Precious metals, etc.
20. Other metal industries

(Unit: million dollars. Data in Table *A*, Appendix A.)

ment of the bars is governed by considerations exactly similar to those which apply in constructing a corresponding summary table.

A second large class of categorical charts includes *statistical maps*. Their use is obviously limited to the presentation of series classified upon a geographical basis. Of the many excellent varieties of statistical maps, the least complicated are the crosshatched map (Charts 9 and 10), the dot map (Charts 11 and 12), and the colored map. The crosshatched type is the most generally

available of the three. It is usually planned so that darker areas represent larger statistical items, but a *legend* giving the numerical significance of each shading must be attached (Chart 9) or the data must be entered on the map (Chart 10). Although the scheme admits of the presentation upon a single map of items widely differing in size, it satisfactorily classifies the items into at most a small number of size groups.

The dot map can generally be used to show items which fall into fairly numerous groups (Chart 11). But it is less satisfactory when the total range in size is very large (Chart 12), because crowding of the numerous dots, influenced by the comparative area of the states, may suggest blackness in the wrong places. Some extension of the range is attained by using dots of differing kinds—dots of different sizes, or circles with shadings or with sectors omitted. Exact geographical location can also be exhibited by a dot map.

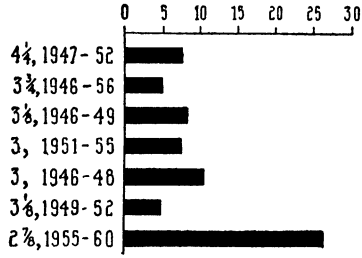


CHART 6.—United States bonds, of over ten years maturity, outstanding on December 31, 1935.
(Unit: hundred million dollars. Data in Table *B*, Appendix A.)

The colored map has substantially the same advantages and disadvantages as the crosshatched map. The excessive cost of reproduction for color plates, however, prevents their extensive use in statistical publications. The colored map is therefore restricted largely to study maps for the investigator and wall maps for lectures and exhibitions.

Another special charting device available for a limited class of categorical series is the *component-part diagram*. The divided rectangle (Charts 13 and 14) and the pie diagram (Chart 15) are of this sort. The obvious purpose is to present the individual items in relation to the total. In the rectangle this can be done by placing end to end successive segments of a rectangle, the length of each segment being proportional (on some selected scale) to the size of the corresponding item, and allowing the total length of the rectangle to be determined automatically (Chart 13). The other plan requires simply the division of a rectangle of given length into segments proportional to the several items, and this necessitates expressing the items as percentages, or other convenient fractions, of the total (Chart 14). This second device has certain advantages if the proportionate importance of the constituents of two compar-

able totals, for example the aggregate income in two different years, is to be related on a single chart.

In constructing the pie diagram, also, the data should first be expressed in the ratio form, and the division of the circle is then accomplished by laying off successive central angles the ratios of

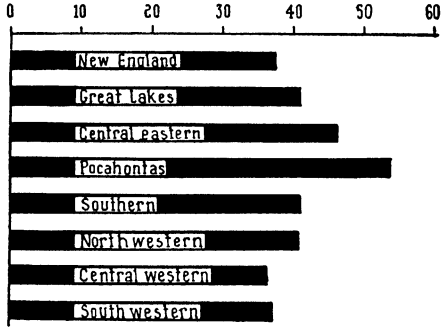


CHART 7.—Percentage of net to gross ton-miles (excluding locomotives) for Class I railroads in the United States in 1936

(Data in Table *C*, Appendix A)

which to 360° are proportional to the ratios computed from the data (Chart 15). Although the circular figure is very widely used, mainly because of its popular appeal and its suggestion of completeness, the rectangular scheme is preferable. The chief

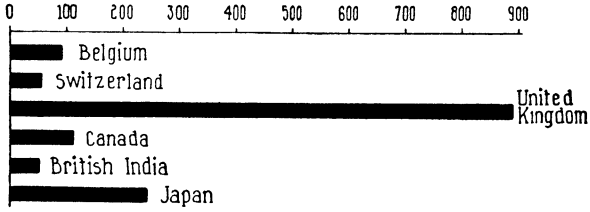


CHART 8 —United States net imports of gold in 1937 from countries making principal shipments.

(Unit: Million dollars. Data in Table *D*, Appendix A.)

reason for this is that the eye obtains its estimate of proportion in the pie diagram by judging the relative angular width of the sectors and in the divided rectangle by judging the relative length of the divisions, and a basic rule of charting is that the property of a figure which can be assessed most easily and most accurately is length. A further disadvantage is that because the various sectors are not similarly oriented on the diagram an optical illusion may result. In all charts of the component-part type an issue

always arises as to the order of arrangement of the elements—segments of a bar or sectors of a circle. The ruling considerations are largely the same as those determining order of items in a summary table.

Mention might be made of a large number of other statistical charts, many of which scarcely merit the name. Many of these use the so-called pictorial feature, which seeks to label and perhaps measure a statistical item by a sketch—such as a ship, or automo-

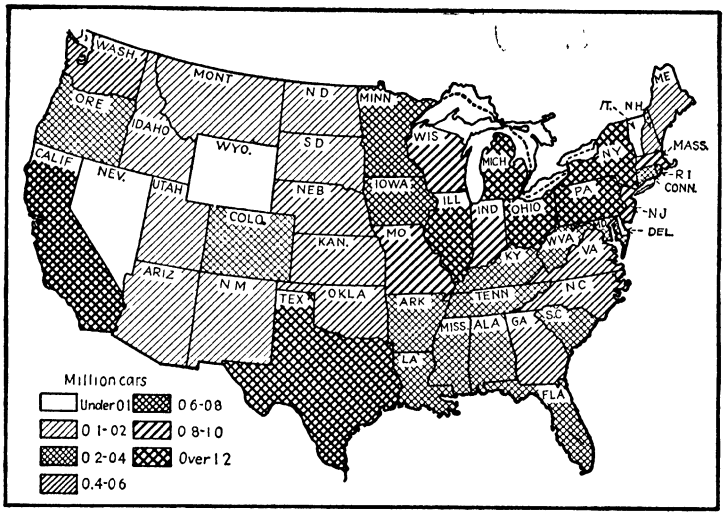


CHART 9.—Classification of states according to number of motor cars registered in 1936.

(Data in Table 25, page 79.)

bile, or man, or bull—of the object enumerated statistically; whatever the advertising value of such tricks, their scientific deficiencies are numerous and serious.

Numerous other examples of complicated diagrams which encumber many popular publications need little comment; with few exceptions they could well be discarded. The mere complexity of some of these schemes renders incorrect inferences, by the analyst as well as the reader, almost inevitable. The student will discover that scientific graphic presentation depends upon simplicity and dignity, and he will do well to avoid the use of special charting schemes which are usually confusing to the reader and often convey impressions positively false.

One of the chief criticisms against the pictorial diagram is that there may be confusion or error in reading size from such charts.

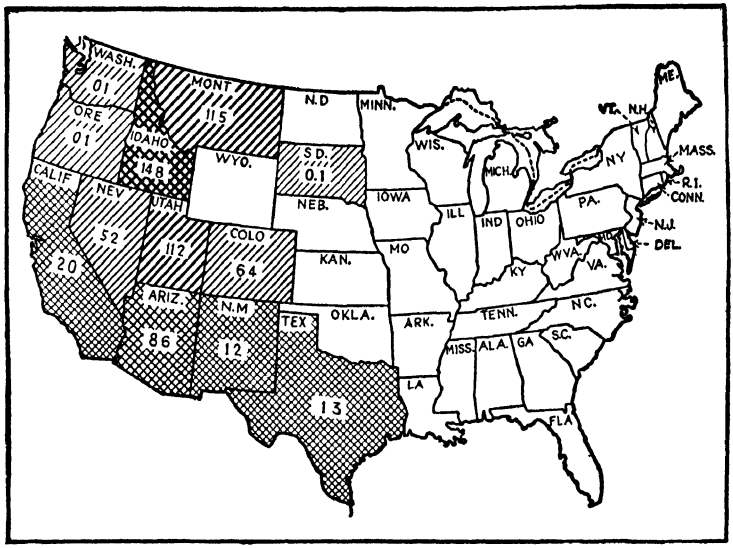


CHART 10.—Silver production in chief producing states in 1936.
(Unit: million fine ounces. Data in Table *I*, Appendix A.)

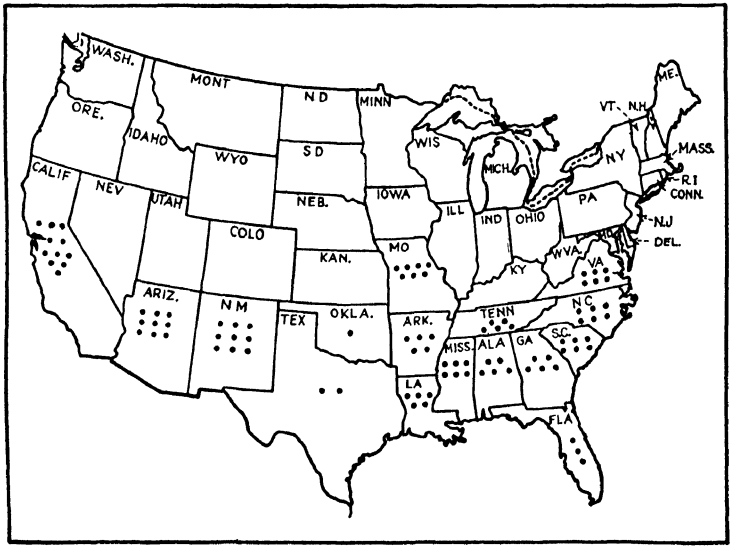


CHART 11.—Yield per acre of cotton in chief producing states in 1936.
(Each dot represents 50 pounds per acre or nearest fraction. Data in Table *O*, Appendix A.)

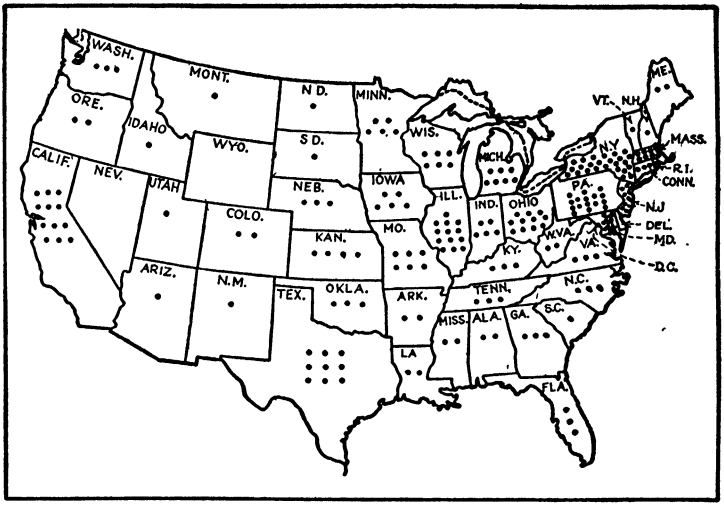


CHART 12.—Number of commercial concerns in each state in 1936.
(Each dot represents 10,000 firms, or nearest fraction. Data in Table *P*, Appendix A.)

Individual income tax	Corporation income and excess profits taxes	Alcohol taxes	Estate and gift taxes	Tobacco manufactures tax	Sales taxes	Other miscellaneous taxes	Social security taxes
1091.7	1082.0	594.2	305.5	552.3	617.4	[illegible]	265.7

CHART 13.—U. S. Internal Revenue collections for fiscal year ending June 30, 1937.
(Unit: million dollars. Data in Table *E*, Appendix A.)

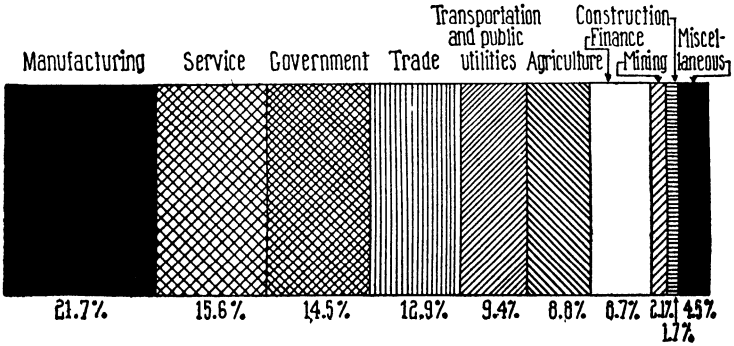


CHART 14.—Percentage distribution of national income of United States in 1935 according to industrial origin.
(Data in Table *F*, Appendix A.)

The principle involved is illustrated in Chart 16, which shows how two items appear, one twice as large as the other, according as the comparison rests upon length, area, or volume. In the complicated diagram also, even though it is not of the pictorial type, the comparisons of length are usually obstructed by the very intricacy of the arrangements. The fundamental rule will bear repetition: Comparisons of *length* are those most promptly and accurately made by the eye.

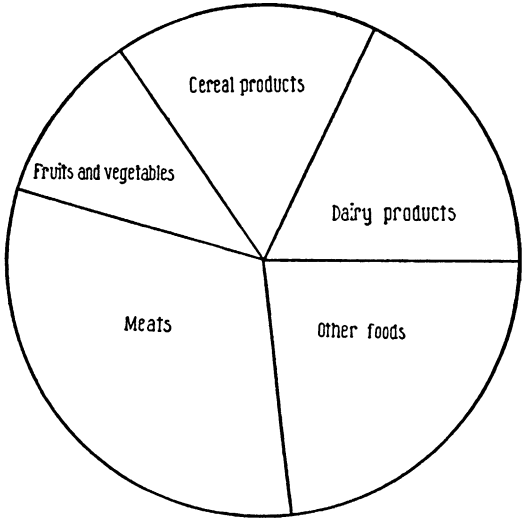


CHART 15 —Percentage weights of subgroups in the foods group of the U. S. Bureau of Labor Statistics wholesale price index.
(Data in Table G, Appendix A.)

Little need be said concerning the more general rules of charting. The same comments on the selection of a title as were set forth in the discussion of tabulation apply to charting. The size of a particular chart will be controlled largely by the items to be included, but the purpose of the chart will play a part in the decision. Charts for publication are generally limited by the size of the printed page, it being understood that the actual drawing is ordinarily made oversize with a view to its reduction in the engraving process. The ratio of height to width is determined partly by consideration of appearance and balance. So far as possible, tabulated data should be presented in connection with the corresponding charts, or reference should be made to an easily accessible source for the data, so that readers may check visual

impressions and also obtain the exact numerical items for those further studies suggested by the chart. As in the case of tabula-

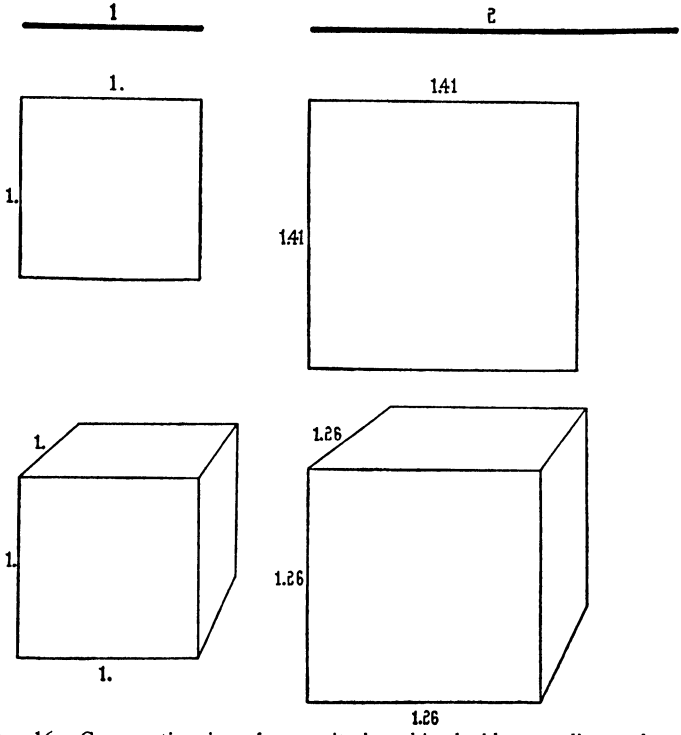


CHART 16.—Comparative sizes of a magnitude and its double, according as they are represented by lines, areas, and volumes.

tion, the making of charts is an art which the student will acquire only by long and critical observation of the work of others and by actual experience with his own graphing problems.

CHAPTER VIII

CHARTING: TIME SERIES

USE OF CHARTS IN ANALYSIS AND EXPOSITION

Charting is generally a preliminary or intermediate step in the analysis of time or frequency series. A limited use of charting can also be made as an aid in the more extended arithmetical analysis of categorical data. Moreover, charts often constitute the final results—the medium of presentation—of statistical analyses of time or frequency series, as well as categorical series. The fact remains, however, that for time or frequency statistics charting is even more important as a tool of research than as a method of presentation. So great is the importance of the *graphic method* in the study of statistical data, both as one of the steps in such study and as a process to be followed concurrently with arithmetical study, that the *analytical* as well as the *expository* point of view will be stressed in the present chapter and in the chapter on frequency charts which follows.

The wide use of time and frequency charts in presenting data and the results of analysis, on the other hand, emphasizes the need for careful attention to the requisites and difficulties of good presentation. In the preparation of charts for presentation, the same basic rules apply in the case of time or frequency data as in the case of categorical data. Simplicity in plan and dignity in execution are always desirable. A clear and adequate title, unmistakable specification of scales, and concurrent tabular presentation of data (or specific reference to an easily accessible tabulation) are essential features of every published chart. In many cases, also, different curves on a single chart require specific labels and distinctive legends.

THE LOCATION OF POINTS WITH ARITHMETIC SCALES

The chief form of chart for time series is the *line graph.* Time data can indeed be represented by a bar diagram, or some special type of chart, but the device which has been found most widely useful is the graph, or curve. The construction of a curve implies the preparation of a *grid.* A horizontal line is chosen as the *time*

axis, and a line perpendicular to it as the *variable axis*. The point of intersection of the two lines is the *origin* and is generally the

TABLE 37

ANNUAL EARNINGS OF FEDERAL RESERVE BANKS ON HOLDINGS OF UNITED STATES GOVERNMENT SECURITIES*

1920	7.14	1929	8 16
1921	6 25	1930	17.27
1922	16 68	1931	12 43
1923	7 44	1932	26 92
1924	14.71	1933	37.53
1925	12 78	1934	46 13
1926	12 59	1935	39.80
1927	14 21	1936	35 18
1928	10 83		

* Unit: million dollars. Source: "Twenty-Third Annual Report of the Board of Governors of the Federal Reserve System," Washington, 1937, p. 98.

zero point for both time and variable scales. If, as is usually the case, there are no negative values of the time or of the variable for the series to be plotted, the origin is chosen near the lower left corner of the chart, and the axes are ruled to the right and upward from the origin. Thus, for the data of Table 37, the first step leads to the rulings shown at the left and bottom of Chart 17. Next, the horizontal (time) scale should be chosen so that (1) the total time interval to be covered will fall within the chart of the size planned and (2) the subdivisions of time can be measured conveniently with available measuring sticks or printed scales. The vertical (variable) scale is determined by similar considerations. Manifestly, the two scales are independent of each other: the number of years per inch horizontally is not dependent upon the number of million dollars per inch vertically. The results, with the earliest figure for the time scale as 1920 rather than zero,

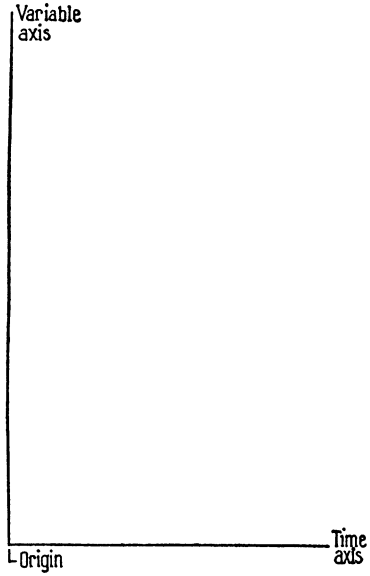


CHART 17.—Selection of axes, for charting the data of Table 37.

appear along the left and bottom edges of Chart 18. By measuring time to the right from the variable axis and the variable upward from the time axis, any particular point (as the point for 1923 in Chart 18) can be located, and similarly all points corresponding to tabulated variates can be plotted. (Actually, the final point of the table is not plotted.) A horizontal measurement of this sort is called an *abscissa* and a vertical measurement an *ordinate*. Both the amount and the direction of these measurements are fixed: measurements horizontally to the right or vertically upward are positive (plus), measurements horizontally to the left or vertically downward are negative (minus). A pair of numbers (each with the appropriate plus or minus sign attached), the first an abscissa and the other an ordinate, definitely fix the location of a point on the chart. Such a pair is called the *coordinates* of the point; and the pair is sometimes used as a label for the point, the coordinates appearing in parentheses, as (3, 7.44).

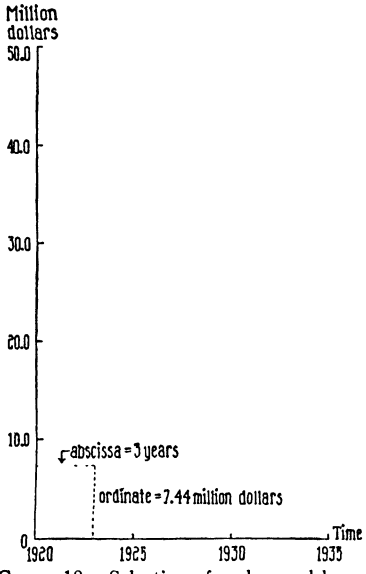


CHART 18.—Selection of scales, and location of one point, for the data of Table 37.

THE CONSTRUCTION OF A GRID

The location of points is much facilitated by the establishment of a *grid*. This consists in ruling horizontal and vertical lines through the principal scale marks on the vertical and horizontal axes, respectively (see Chart 19). By means of grid lines, one can readily and with fair accuracy locate the approximate position of the several points. The vertical grid lines can ordinarily be selected so that plotted points will fall exactly on the lines (see Chart 20), but are sometimes located so that the position, so far as the time measurement is concerned, must be estimated (see Chart 21). The horizontal grid lines can seldom be chosen so that plotted points will fall upon them; the vertical measurement must

usually be estimated by aid of the grid lines, or measured with a movable stick or printed scale.

Unless a ratio scale (see page 112) is used, all horizontal grid lines should be equally spaced, and likewise all vertical lines. In the construction of the working chart this rule can be relaxed, but in the complete chart, whether for presentation or for future study, the spacing should—in most cases, must—be uniform. Otherwise

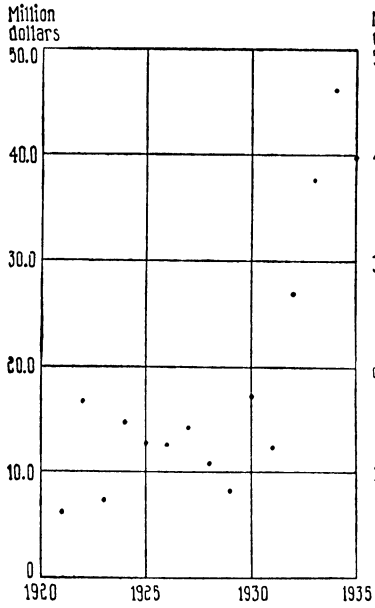


CHART 19.—Annual earnings of Federal Reserve Banks on holdings of U. S. Government securities.
(Data in Table 37, page 103.)

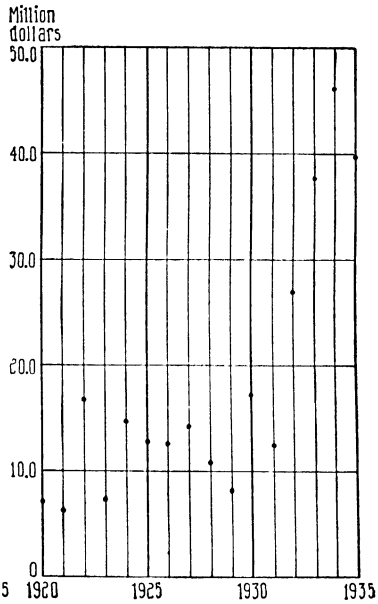


CHART 20—Supplementary vertical grid lines used in plotting Chart 19, but not intended for presentation.

as it is customary to assume that the space between every pair of adjacent rulings represents the same amount of change in the variable, we are likely to make misleading inferences in reading the chart. Chart 22 illustrates the rulings used in plotting the points (broken lines), and those which are to be kept in the finished figure (solid lines).

In preparing charts for presentation, the number of grid lines retained is controlled by a balance between two conflicting points of view: ease and speed in reading the position of specific points, and subordination of the grid to the curve itself. The grid rulings should never be heavy, and they should be spaced so

widely that the final chart appears "clean." The top of the curve should ordinarily fall well within the grid (Chart 31 rather than Chart 24), but an exceptional point may run to the top of, or even above, the grid (see Chart 25).

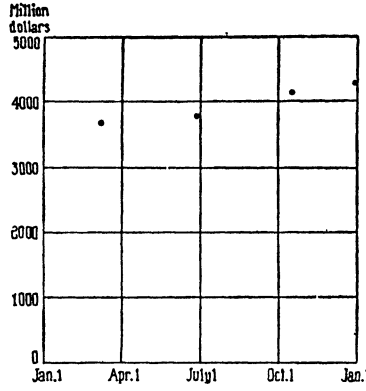


CHART 21.—Adjusted demand deposits of all member banks of Federal Reserve System at dates of call in 1934.
(Data in Table *H*, Appendix A)

The height of the grid should normally be somewhat less than the width (a ratio of 3:5 or 2:3 yields the most pleasing appearance), but this requirement must be tempered somewhat by the shape of the page upon which the chart is to be presented, as well as by the nature of the data. (In Charts 17 to 20 this rule was not observed.) In case the variable does not fall nearly to zero (Table 38), the scale can be expanded within the available space by omitting a section of the vertical scale, but this omission should be marked by an interruption in the grid rulings, lest the reader base his judgment upon the apparent position of the curve with reference to zero (Chart 26). An unused portion of the horizontal scale can also be deleted in a similar manner if conditions warrant (Chart 27).

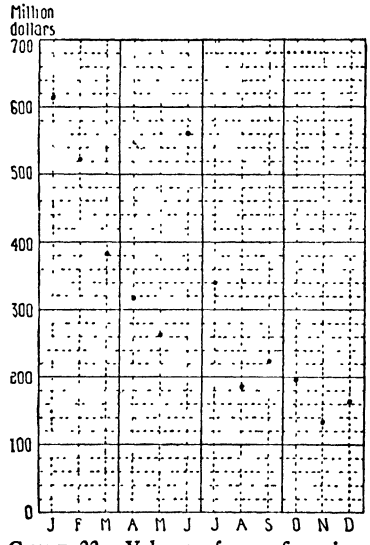


CHART 22.—Volume of new financing in the United States in 1937.
(Data in column *a* of Table *J*, Appendix A.)

The variable may extend below the time axis (Chart 28); and the range of variation of time often includes negative numbers (Chart 29), if for some reason a particular time within the interval covered is chosen as zero. These negative values of the time and of the variable seldom occur in the original data but they fre-

quently arise in the process of analysis, and charting devices for presenting them are essential. In some special charts the position of the time axis does not correspond with the zero of the vertical scale (Chart 30), but this is a formal rather than a real deviation from the general rule. In such cases, the bottom "ruling" should be irregular rather than straight, to call attention to the omission of zero.

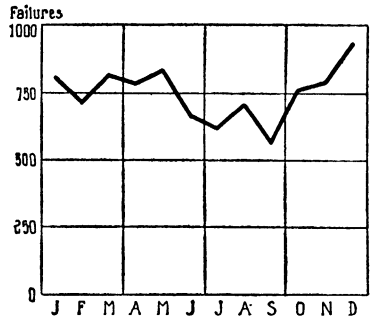


CHART 23.—Number of business failures in 1937.
(Data in column *b*, Table *J*, Appendix A.)

The ruling of grids for the plotting of the points is a laborious process and is usually avoided by the use of prepared forms. Ruled forms are printed or lithographed and obtainable in considerable variety, the differences pertaining chiefly to the spacing of horizontal or vertical rulings, the relative weight and color of the lines, and the size of sheet. Here, as in other phases of practical statistics, too great diversification is enormously wasteful. A small number of simple forms, covering the cases most commonly met in practice, should be sufficient. (Specific suggestions are given in Appendix B.) Forms with rulings of certain colors, particularly blue, are very convenient for photographic reproduction: the detailed blue grid lines used in plotting the points disappear in the photograph, and the black grid lines, ruled by hand only at the positions where it is desired that they appear in the cut, are reproduced.

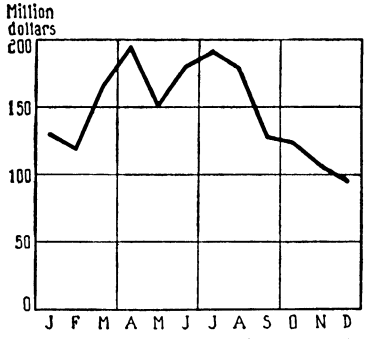


CHART 24.—Value of construction contracts awarded, and privately financed, 1937.
(Data in column *c*, Table *J*, Appendix A.)

Ruled forms can be used also as *backgrounds* in plotting curves upon tracing cloth or translucent paper. Particularly for the making of study charts, this plan is economical and has great flexibility. The cloth or paper is placed over the background, and only those grid lines which are to be retained are ruled upon the

chart. This ruling of grids, and the marking of the corresponding scales, should probably be done before plotting points, lest the tracing be displaced with consequent inaccuracy. The preparation of a moderate number of backgrounds provides a sufficient variety of scales, as different scales for different charts can be assigned to the rulings of a single background. Moreover, by using a translucent chart form with fixed vertical rulings over a fixed background with closely spaced horizontal rulings, a large number of charts of one type with different vertical scales can be prepared.

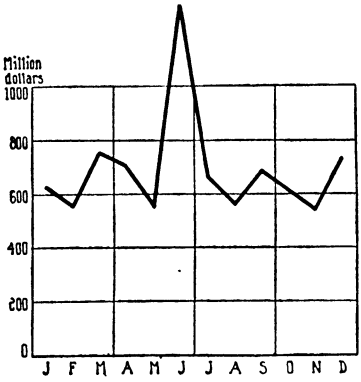


CHART 25.—Total expenditures of the U. S. Treasury in 1937.
(Data in column *d*, Table *J*, Appendix A.)

ACTUAL AND PERCENTAGE VARIATION

The discussion above assumes that the vertical scale is *arithmetic*, that equal *amounts* of change in the variable are represented by equal vertical distances on the chart. There is frequent occasion, in plotting time series, to emphasize comparisons of rates of change rather than amounts. For this purpose the *ratio scale*

TABLE 38
MAIL ORDER AND STORE SALES, TWO COMPANIES, MONTHLY, 1937*

January	54 4	July	73.7
February	53 8	August	71.3
March	78 6	September	90 2
April	89 7	October	107.5
May	92 6	November	89 8
June	89 3	December	116 2

* Unit: million dollars. Source: *Survey of Current Business*, February, 1938, p. 27.

(logarithmic scale) is most useful. This scale is designed so that equal *percentages* (rates) of increase are represented by equal vertical distances. Whereas with an arithmetic scale the rise on a chart from 40 to 48 would appear as 4 times the rise from 10 to 12, with a ratio scale the two rises would be equal. The solid line curve of Chart 31, plotted from the items in column 1 of Table 39, presents diagrammatically the fluctuations in the actual security issues. From this curve it is possible to estimate the *amount of change* in issues between any pair of months, to gain a satisfactory impres-

sion of the direction in which security issues moved at various times, and to ascertain the times when changes in the direction of

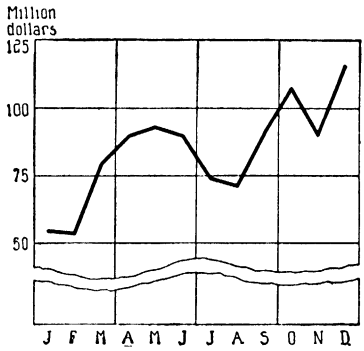


CHART 26.—Sales of two mail-order companies in 1937.
(Data in Table 38, page 108)

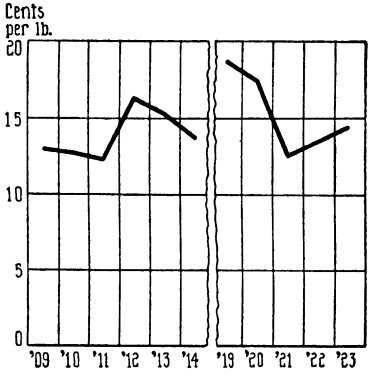


CHART 27.—Average yearly price of electrolytic copper at New York.
(Data in 1925 ed. of "Economic Statistics," page 394.)

this movement occurred. The relative amount of change in issues cannot, however, be assessed easily: to determine whether the *rate of increase* in issues at one time was greater or less than that at another time is not easy from this solid curve. Thus, examination of the solid curve indicates that the actual increase in issues from March to April, 1936, was greater than that from August to September, 1936, and that the actual decline in issues from December, 1936, to January, 1937, was greater than that from August to September, 1937. But it does not show that the comparative *rate* of change in each of these cases was precisely opposite to the comparative *amount* of change. In fact, as the figures of column 3 show, the percentage change from March to April, 1936, was less than from August to Sep-

CHART 28.—Statutory net income less deficit of all corporations in the United States.
(Data in Table K, Appendix A.)

tember, 1936, and likewise that from December, 1936, to January, 1937, was less than that from August to September, 1937.

It is therefore desirable to construct a chart in such manner as to set forth accurately the comparison between rates of change in various years. Column 4 of Table 39 gives the logarithms of the

TABLE 39

TOTAL REFUNDING ISSUES OF CAPITAL IN THE UNITED STATES, MONTHLY, 1936–1937*

Year and month	Amount[a] (1)	Actual change[a] (2)	Per cent change (3)	Logarithm of amount (4)	Item in col. (1) ÷ 3^x (5)
1936					
January	287.6	.....	.....	2.46	95 9
February	195.8	− 91 8	− 32	2 29	65 3
March	638.4	+442 6	+226	2.81	212.8
April	826 9	+188.5	+ 30	2.92	275 6
May	308 3	−518.6	− 63	2 49	102 8
June	515.7	+207.4	+ 67	2.71	171.9
July	235.6	−280.1	− 54	2.37	78 5
August	80.4	−155 2	− 66	1.91	26.8
September	231.5	+151.1	+188	2.36	77 2
October	277 5	+ 46 0	+ 20	2.44	92.5
November	222 9	− 54 6	− 20	2 35	74.3
December	459.4	+236.5	+106	2.66	153.1
1937					
January	374 4	− 85 0	− 19	2 57	124.8
February	354 2	− 20 2	− 5	2.55	118.1
March	197.0	−157 2	− 44	2.29	65 7
April	158 2	− 38 8	− 20	2 20	52 7
May	116.3	− 41 9	− 26	2.07	38.8
June	200.5	+ 84.2	+ 72	2.30	66.8
July	93.4	−107 1	− 53	1 97	31 1
August	108.6	+ 15 2	+ 16	2 04	36 2
September	66 8	− 41.8	− 38	1 82	22 3
October	107 0	+ 40.2	+ 60	2.03	35.7
November	41 5	− 65.5	− 61	1 62	13 8
December	42.1	+ 0.6	+ 1	1.62	14 0

* Source: *Survey of Current Business*, February, 1938, p. 21.
[a] Unit: million dollars.

items of column 1, and the dotted curve of Chart 31 rests upon these logarithms as data. This curve gives a true picture of the percentage changes mentioned above. The reason for this is obvious from a principle of logarithms: the difference between the logarithms of a pair of numbers is greater than the difference between the logarithms of another pair of numbers if the ratio between the numbers of the first pair is greater than the corre-

sponding ratio between the numbers of the second pair. Thus the ratio of the amount of issues for April, 1936, to that for March,

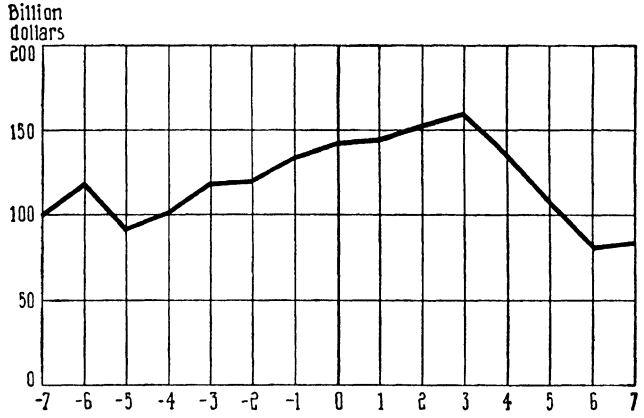


CHART 29.—Gross income of all corporations in the United States, 1919–1933, centered at 1926.

(Data in Table L, Appendix A.)

1936, is less than the ratio of the amount of issues for September, 1936, to that for August, 1936, and the difference in the logarithms for the first case is 0.1123, whereas it is 0.4592 for the second case.

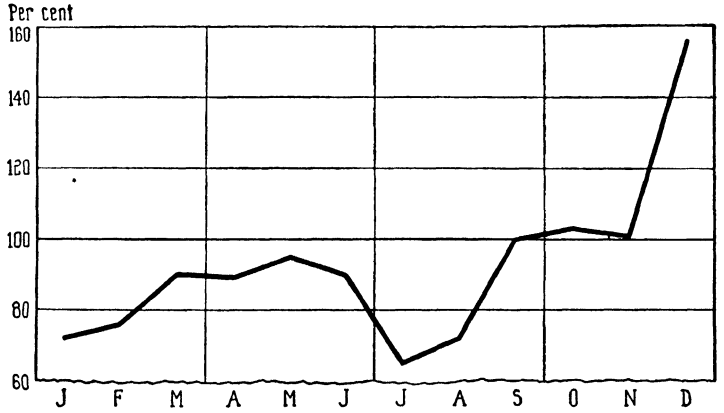


CHART 30.—Actual department store sales, 1937, expressed as per cent of 1923–1925 average.

(Data in column *e*, Table J, Appendix A.)

It is the appraisal of these two differences, 0.1123 and 0.4592, which the eye makes in studying the dotted curve. A similar

explanation holds for the other comparison—percentage change from December, 1936, to January, 1937, with percentage change from August to September, 1937—if it is borne in mind that these changes are declines and that differences of the logarithms are therefore negative. This dotted curve is then a device for graphically exhibiting percentage changes, and this method of charting is the one ordinarily desirable in a graphical study in which the interest centers upon rates of change.

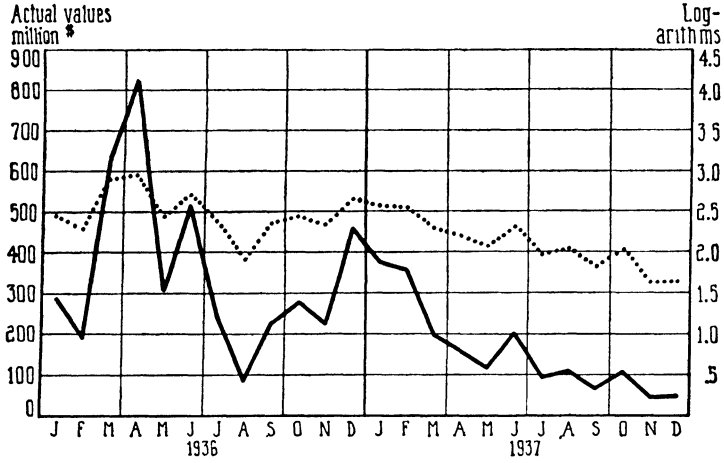


CHART 31.—Total refunding issues of capital in the United States, 1936–1937, actual figures, solid curve, their logarithms, dotted curve.

(Data in Table 39, page 110)

THE RATIO SCALE

The looking up of the logarithms is, however, an obstacle to the use of this sort of chart. That step in the tabulation can be avoided by the use of paper ruled on a *logarithmic* or *ratio scale.* Such a set of rulings is used in Chart 32. The principle on which these rulings are made is the following: A horizontal line is chosen and marked 100, and the logarithm which corresponds to it is 2.000; and next the logarithms are found from a logarithm table for the numbers 30, 60 . . . , 600, 800, and any intermediate numbers desired; and, using a convenient scale in inches and making measurements from the chosen 100 line (log 100 = 2.000), the other lines of the chart are located. Thus, the distance from the 100 line to the 200 line is proportional to the difference between their logarithms, 2.000 and 2.301; and the distance between the

100 line and the 300 line is proportional to the difference between their logarithms, 2.000 and 2.477. Thus, having selected the 100 line and having chosen a scale (say 1 inch to a unit of logarithms) the 200 line is drawn 0.30 inch above the 100 line and the 300 line is placed 0.48 inch above the 100 line. In this way, all the horizontal lines can be located.

A striking feature of the completed rulings is that spaces increase in width toward the bottom, and indeed it will be found impossible to locate any line to correspond to the number zero,

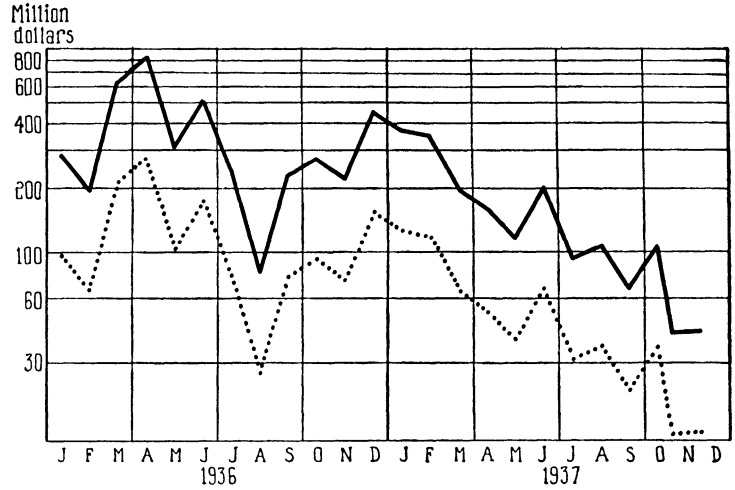


CHART 32.—Comparison, on logarithmic scale, of actual data (solid curve) of Table 39 and Chart 31, with those data divided by 3 (dotted curve).

which has no finite logarithm. It is for this reason that a chart on the ratio scale never has a zero line. The widening space at the lower end of the scale suggests the wisdom of introducing intermediate rulings on that part of the scale. Generally the horizontal rulings, unlike those of Chart 32, should be equally spaced. Thus those of Chart 32 might advantageously have run: 25, 50, 100, 200, 400, and 800—all spaces would then have been equal.

Moreover, it is not necessary to attach specific numbers to the rulings of a ratio chart, although such scale marks are generally useful in completed charts. Thus the same set of rulings would be valid for the numbers 6, 8, . . . , 20, . . . , 38, 40, as for 3, 4, . . . , 10, . . . , 19, 20. The essential point is that the ratio scale exhibits relations between the *ratios* of numbers and not

between their actual values. This brings out another feature of the ratio scale: once the rulings are made, the rate of change per vertical inch (or other linear unit) is fixed. Thus for a particular set of rulings a vertical rise of 1 inch anywhere on the chart represents a fixed percentage increase in the variable. A vertical drop of 1 inch, however, would represent a somewhat smaller percentage *decrease* in the variable. Even if all the data are multiplied by a constant, the result of plotting on the given ratio scale will be a curve of the same shape as before: the two curves are identical in inclination of lines and range of fluctuation. Chart 32 illustrates this fact: the data of column 1 in Table 39, and the same data divided by 3 plotted on the same ratio scale, in the solid and dotted curves, respectively.

In order to change the resulting curves, the scale of the rulings must be changed. A ratio scale on which the variable increases in the ratio of 1:10 upon a vertical movement of 5 inches is called a *5-inch scale.* Likewise there are 10-inch scales, 20-inch scales, and, in fact, a ratio scale can be constructed for any specified rate of increase per inch. One can conveniently construct a 10-inch scale by using the graduations on the lower edge of the slide of a 10-inch slide rule, although the exact length of the range from 1 to 10 on a slide rule is often not precisely 10 inches. The fundamental method of constructing a ratio scale is, of course, by means of logarithms, as described above. (See also Appendix B, for suggestions about constructing scales.)

INTERPOLATION AND SMOOTHING

The plotted points of a graph are joined ordinarily by a smooth curve or succession of line segments. The justification for this lies in the use of the continuous line either as an aid to the eye in following the plotted points or as a means of estimating values of the variable for times other than those specified in the data. The second reason clearly does not apply in cases, such as Chart 33, in which the points actually plotted pertain to totals or averages for intervals of time. Even where the plotted points pertain to instants of time, as in Chart 34, the use of a line segment joining two successive points as a basis of estimating values of the variable between the two given instants is open to objection. This is especially the case if, as in Charts 35 and 36, the actual fluctuation shown by the plotted points is wide or highly irregular.

This method of using line segments for estimating the value of the variable at a point between the points actually plotted is called

interpolation.[1] The corresponding operation of *extrapolation*, by which a value of the variable is estimated for a time beyond the total time interval covered by the plotted points, is even more hazardous. Although the customary practice in making graphs is to join the plotted points by a succession of line segments, it should be understood that in relatively few cases may the resulting "curve" be used as a proper interpolating or extrapolating device. Usually the curve on a time-series chart is merely an aid to the eye in tracing the fluctuations indicated by the plotted points.

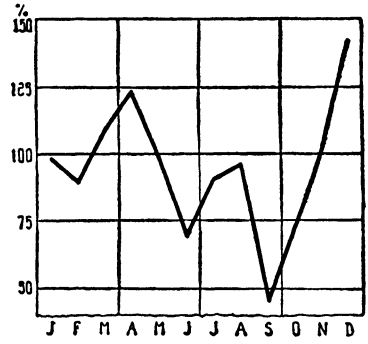


CHART 33.—Sugar 'meltings' in 1937, expressed as percentages of 1923–1925 average and adjusted for seasonal variation.

(Data in column *f*, Table *J*, Appendix A.)

The *smoothing* of a curve is the removing from the line graph of those irregularities, or supposed irregularities, which are indicated by sharp changes in direction of the line segments. Smoothing can be carried out by inspection or by mathematical adjustment. The former

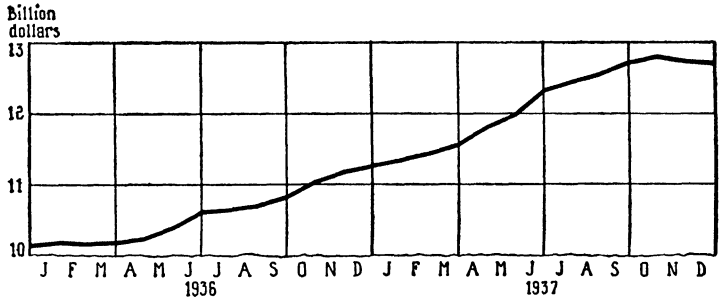


CHART 34.—Monetary gold stock of the United States at the end of each month, 1936–1937.

(Data in column *g*, Table *J*, Appendix A.)

method, by far the more commonly used, requires skill and experience; whereas the latter method, which will be mentioned more fully below (Chap. XIII), rests upon certain fundamental

[1] Interpolation is possible also by the use of more complicated curve segments than the straight line, but assumptions are implied, in the selection of any particular curve (such as the parabola or compound-interest curve) which require careful examination before important conclusions may be drawn.

assumptions concerning the nature of the fluctuations or the causes from which they arise.

The process of smoothing by inspection is very widely used. It consists in drawing a smooth curve, one in which changes in direction occur gradually, among the plotted points in such manner as to exhibit the significant fluctuations and conceal the irregularities. Obviously, this operation requires judgment of a high order to decide which fluctuations are significant and which are irregularities, and the result can never be entirely free of a subjective element. The large part played by the discretion of the statistician in smoothing is ample reason for urging that the process be used sparingly. Sometimes, to minimize the subjective element, several individuals make independent smoothings of the same set of plotted points; the several results are compared, and summarized in a single compromise curve. Chart 37 is a case in which smoothing is helpful, although the resulting "fit" of the curve is not close.

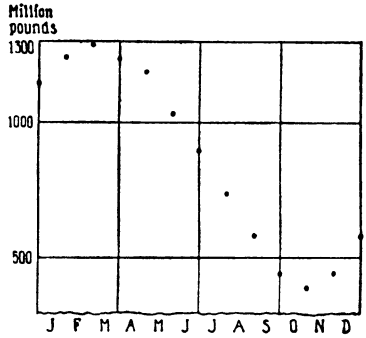


CHART 35.—Stocks of meats in cold storage, end of each month in 1937.
(Data in column *h*, Table *J*, Appendix A.)

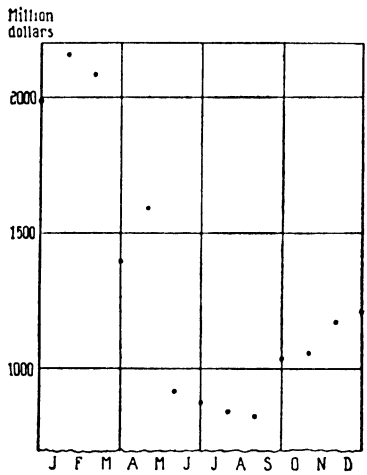


CHART 36.—Excess reserves of member banks of Federal Reserve system near end of each month in 1937.
(Data in Table *M*, Appendix A.)

COMPARISON OF TIME GRAPHS

The comparison of two or more series by graphic means is a very effective aid to statistical analysis of time series. If several curves are plotted upon the same chart, with identical time scales and identical or appropriately related vertical scales, the reader can frequently deduce important relations from a study of the

fluctuations. Comparisons of the amounts of change in different series can best be made upon an arithmetic scale, whereas comparisons of the relative changes—either on one or on each of the curves, or as between the curves—are facilitated by ratio charting.

A peculiar advantage of the ratio chart is that one curve can be shifted vertically relative to the other so that any desired point of one can be brought into coincidence with the corresponding point of the other, without in any way invalidating the comparison—this

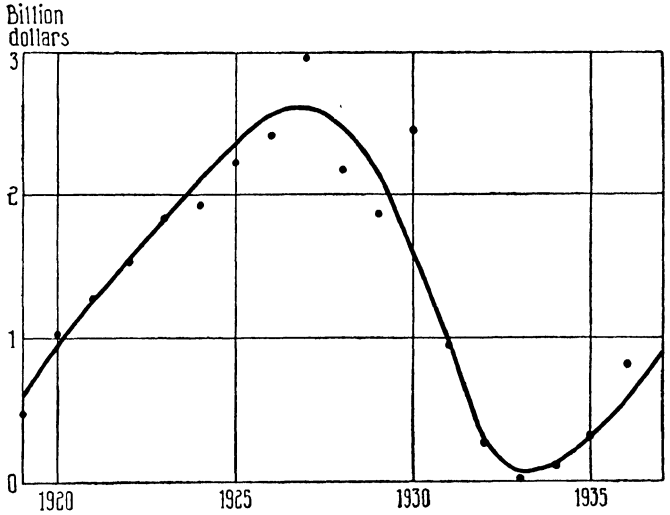


CHART 37.—Issues of corporate long-term bonds, for new capital, in the United States, 1919–1937; with visually fitted curve.
(Data in Table *N*, Appendix A.)

principle is illustrated in Chart 32. This follows from the fact mentioned above, that a given vertical displacement represents the same percentage change everywhere on a ratio chart. This shifting is not possible on the arithmetic scale, for there the zero points (or other proper "normal" points, such as 100 per cent) must coincide (Chart 38). *Different* arithmetic scales, having an identical zero may, however, be used in such cases (Chart 39).

The comparison of more than three or four curves upon a single chart is usually impossible. Except to emphasize great diversity of movement, the placing of many curves on a single chart is futile. Even the practiced eye cannot effectively trace the movements of many individual curves and their relations to one another. For the analysis of several series the process of comparison is aided

by plotting upon translucent paper. Each curve can be plotted in a distinctive color or legend on a single sheet, and comparisons can be made by placing one sheet above the other. Here sheets with curves plotted on the ratio scale can be shifted vertically relative to each other, but charts plotted on the arithmetic scale must be superposed in such manner that the horizontal axes coincide. The horizontal time scales must, of course, be the same for all curves to be compared by superposition.

No attempt will be made in this chapter to discuss the large number of specialized charting devices available for presenting time series. Reference should be made, for information on these types of graphs, to the standard manuals on graphic methods. A word of caution is ventured: Any complicated charting device should be avoided unless there is impelling reason to the contrary, for simplicity is one of the most desirable properties of all graphic presentation.

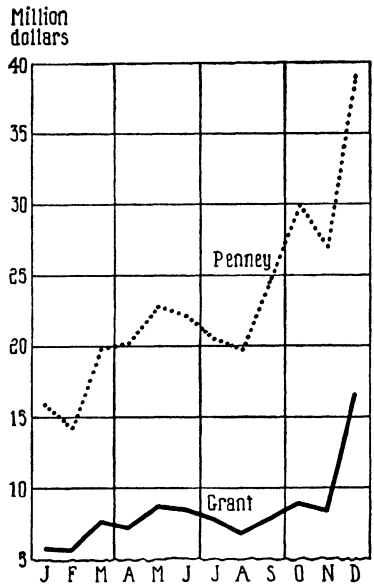


CHART 38.—Sales of two chain-store systems, 1937.
(Data in column *i*, Table *J*, Appendix A.)

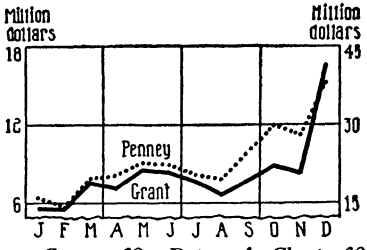
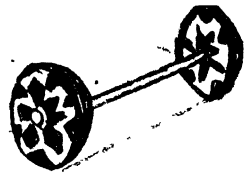


CHART 39.—Data of Chart 38 plotted with two different arithmetic scales (Grant on left, Penney on right) having an identical zero.



CHAPTER IX

CHARTING: FREQUENCY SERIES

THE FREQUENCY POLYGON AND ITS MAIN FEATURES

A frequency series has been defined above (Chap. I) as one in which the rule of classification is a succession of specific sizes, or size intervals, of the variable magnitude in question and each item gives the number of cases or objects for which the variable magni-

TABLE 40

DISTRIBUTION OF BOOKKEEPERS—MACHINE OPERATORS—IN NEW YORK CITY IN MAY, 1937 ACCORDING TO WEEKLY EARNINGS*

Weekly earnings[a]	Number of workers	Cumulative number		Per cent of total
		Below specified limit	Above specified limit	
(1)	(2)	(3)	(4)	(5)
15	88	0	926	9.50
20	229	88	838	24.73
25	372	317	609	40 17
30	139	689	237	15 01
35	54	828	98	5 83
40	14	882	44	1 51
45	10	896	30	1 08
50	15	906	20	1.62
55	1	921	5	0 11
60	2	922	4	0.22
65	1	924	2	0.11
70	1	925	1	0.11
75		926	0	
Total	926			100.00

* Unit: one dollar. Source: *Monthly Labor Review*, January, 1938, p. 215.
[a] Lower limit of class interval.

tude has the specified size, or falls in the specified size interval. A particular value, or size, of the variable magnitude is called a *variate*, and the statistical item (number of cases) is called the *frequency*. The size interval is a range of the variable magnitude between two specified sizes called *class limits*, and is called a *class*

interval. The value of the variable magnitude at the middle of a *class interval* is called a *class mark.* Although the items of a frequency series "vary" in the sense that they differ among themselves, the *variable* is actually the magnitude for which sizes are specified by the several class marks. Thus, in Table 40 (column 1), the variable is the weekly earnings. The situation is quite different (as noted above, page 19) in the categorical series, for which the variable is the item itself. In Table 40, 30–35 (upper limit excluded) is a class interval, 30 and 35 are the class limits of that interval, and 32.5 is the class mark of that interval.

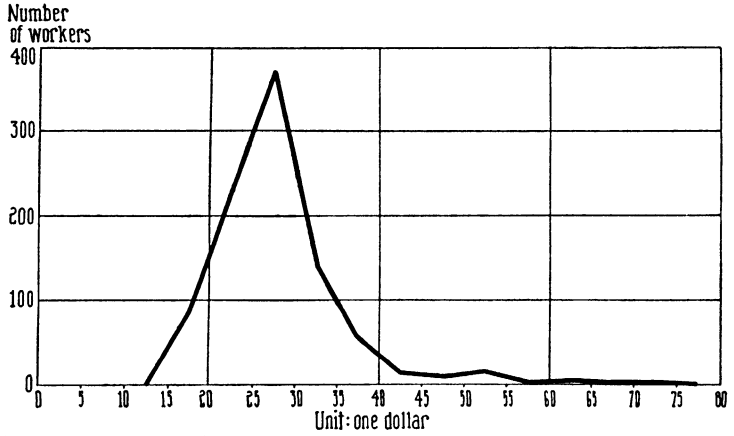


CHART 40.—Number of bookkeepers, machine operators, having weekly earnings in specified five-dollar intervals in New York City in May, 1937.
(Data in Table 40, page 119.)

A common graphic method of presenting frequency series is by means of a *frequency polygon* (Chart 40). A point is plotted for each item of the series by taking the corresponding class mark as abscissa and the item itself (frequency) as ordinate. Each class interval for which there is no item, including the interval next below the lowest interval having a frequency (the interval 10–15 in the illustration) and that next above the highest interval having a frequency (the interval 75–80), are taken as having frequency zero. The plotted points are joined by line segments to yield the completed polygon.

Certain inferences concerning the nature of the plotted series can be drawn directly by study of the polygon. These inferences will be discussed in turn as the several points involved are treated in the elaboration of numerical methods of analysis, but their

nature can be indicated by the following considerations. In a frequency polygon, such as that of Chart 40, the line joining two plotted points does not yield, except in particularly favorable cases, even a tolerably good basis of interpolating between the plotted points. Each plotted point represents the total frequency within a particular class interval; and we do not know, unless data are available for a finer size classification of the variable than that used in plotting, how that frequency is distributed *within* the specific class interval. Thus, for interval 30–35 of Chart 40, the total frequency is 139; but we do not know what portion of it belongs to the subinterval 30–31 or any other subinterval.

We might attempt to estimate the frequency in the subinterval by interpolation from the polygon as plotted: reading from the line joining the plotted point for 25–30 with that for 30–35 yields an estimated ordinate of 240. This obviously cannot be the frequency in the subinterval 30–31, because 240 exceeds the entire frequency, 139, of the interval 30–35. The obvious allowance consists in dividing 240 by 5, as the subinterval has only $\frac{1}{5}$ the width of the entire interval; and this yields 48 as the estimate for the subinterval 30–31. But even this is not dependable: the fact that the polygon has so considerable a break in direction between 30 and 35 would mean that estimates made by this process of interpolation would be much higher left than right of 32.5. To some extent this comparative concentration of frequency in the subintervals left of 32.5 is almost surely valid. But what the interpolation process really assumes is that the distribution of the 139 cases among subintervals is proportional to the several ordinates erected, from the various points between 30 and 35, to the broken line. This assumption may be seriously in error, particularly if the break in the line is considerable. Moreover, we should find the sum of the frequencies allocated to all specific subintervals between 30 and 35 would not be identical with 139. Where the break in the line renders the polygon concave upward, as for the 30–35 interval, the sum would exceed 139; if the polygon were convex upward, the sum would fall short of 139. (Further comment on these points appear below, page 130, in connection with smoothing.)

The representation (see below, page 128) of the series of Table 40 by a *block diagram* (Chart 41), rather than a polygon, suggests a quite different assumption about the distribution of frequency within an interval. Interpolation from such a figure evidently yields $\frac{1}{5}$ of 139 as the estimated frequency for *any* subinterval one

dollar wide between 30 and 35. This allocation of frequency is likewise almost surely in error, but it does not have the second

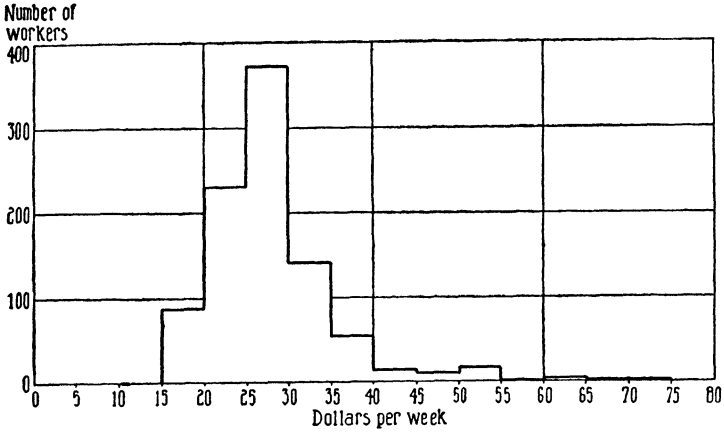


CHART 41.—Block diagram corresponding to the frequency polygon of Chart 40

disadvantage of the polygon case: here the subinterval frequencies do add up to 139. We shall find below that various assumptions

TABLE 41

DISTRIBUTION OF CLERICAL WORKERS, FOR SPECIFIED OCCUPATIONS IN NEW YORK CITY IN MAY, 1937 ACCORDING TO WEEKLY EARNINGS*

Weekly earnings[a]	Bookkeepers' machine		File clerks	
	Number	Per cent	Number	Per cent
10	.	.....	32	1 68
15	88	9.50	751	39.44
20	229	24.73	533	27 99
25	372	40 17	267	14 02
30	139	15.01	195	10.24
35	54	5.83	68	3 58
40	14	1.51	36	1 89
45	10	1.08	13	0 68
50	15	1 62	5	0.26
55	1	0.11	3	0.16
60	2	0.22	1	0.05
65	1	0.11		
70	1	0.11		

* Source: *Monthly Labor Review*, January, 1938, p. 215.
[a] Lower limit of class interval. Unit: one dollar.

need to be made concerning the distribution of frequency within a class interval.

The comparison of two or more frequency series by plotting them upon a single chart is generally less satisfactory than was found to be the case for time series. One of the difficulties is that the total frequency differs for different series, with the result shown in Chart 42 (based upon columns 1 and 3 of Table 41). This obstacle to effective comparison can be removed by expressing

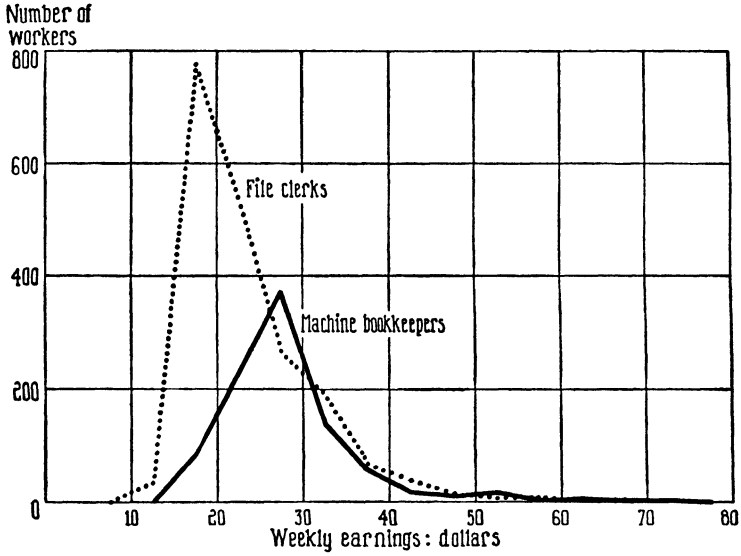


CHART 42.—Frequency polygons comparing earnings of two occupations, among New York City office workers, in May, 1937.
(Data in Table 41, page 122.)

each frequency as a percentage of the total, and plotting the items thus modified (Chart 43, based upon columns 2 and 4 of Table 41). Such percentages of total frequency are called *relative frequencies.* Now another difficulty becomes apparent: the bulk of the frequency pertains to quite different values of the variable for the two series. The difficulty appears in a somewhat different form in Chart 44, the two curves of which are based upon Table 42. The bulk of the frequency falls much nearer the zero of the horizontal scale for female than for male workers, and the spread is much wider for male than female workers.

Closer coincidence could be obtained by shifting the horizontal scale for female workers to the right relative to that for male

workers. Such relative shifting of arithmetic scales is not justifiable in charts for presentation, and must be used sparingly

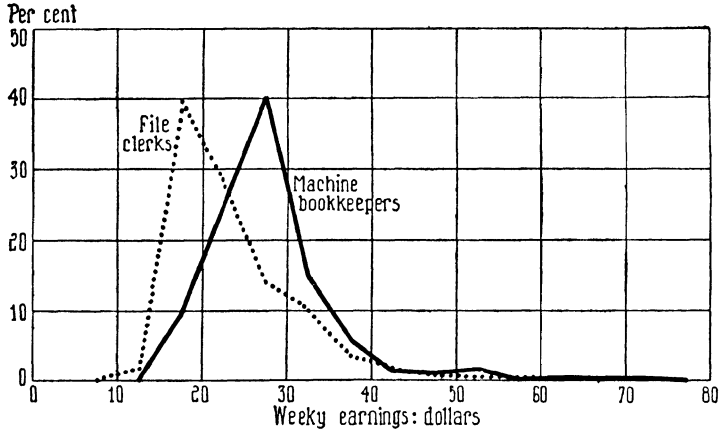


CHART 43.—Relative frequencies for the two occupations, compared in absolute frequencies in Chart 42

even in study charts. The horizontal zeros for both curves should be identical, just as the vertical zeros for two time curves (see

TABLE 42

PER CENT OF EMPLOYEES—HAND AND MACHINE TREERS—IN BOOT AND SHOE INDUSTRY, FINISHING DEPARTMENT, HAVING HOURLY EARNINGS IN SPECIFIED RANGES IN 1930*

Hourly earnings[a]	Male	Female
10– 20	[b]	1
20– 30	6	22
30– 40	10	39
40– 50	23	28
50– 60	21	7
60– 70	19	2
70– 80	12	1
80– 90	6	
90–100	2	
100–110	1	

* Source: "Wages and Hours of Labor in the Boot and Shoe Industry, 1910 to 1930," Washington, U. S Bureau of Labor Statistics, Bulletin 551, February, 1932, p. 27. (Data as given in source have been adapted to intervals shown here. In particular, highest frequency in male series has been *assumed* between 100 and 110 whereas recorded between 100 and 120)

[a] Upper limit of class interval excluded. Unit: one cent.

[b] Less than 1 per cent.

preceding chapter). If the horizontal scale were a ratio scale this objection would be removed. The difference in spread could be

eliminated largely in the present instance by expanding the horizontal scale for females relative to that for males; but such intricate adaptations of scales are of doubtful reliability, even for working

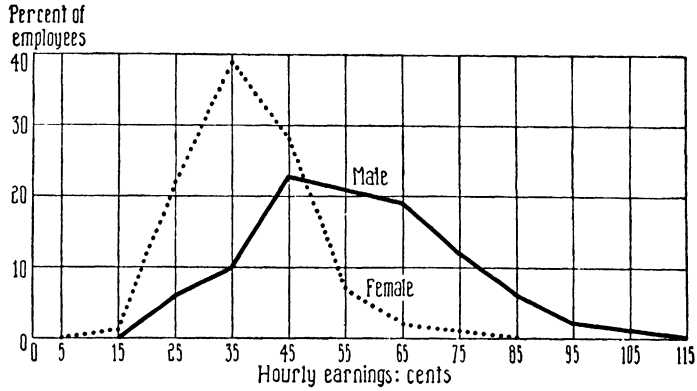


CHART 44.—Polygons comparing relative frequencies of male and female employees—hand and machine treers—in the boot and shoe industry, according to hourly earnings in 1930.

(Data in Table 42, page 124.)

charts. Chart 45 compares the curves of Chart 44, with the female scale shifted to the right; and Chart 46 compares them,

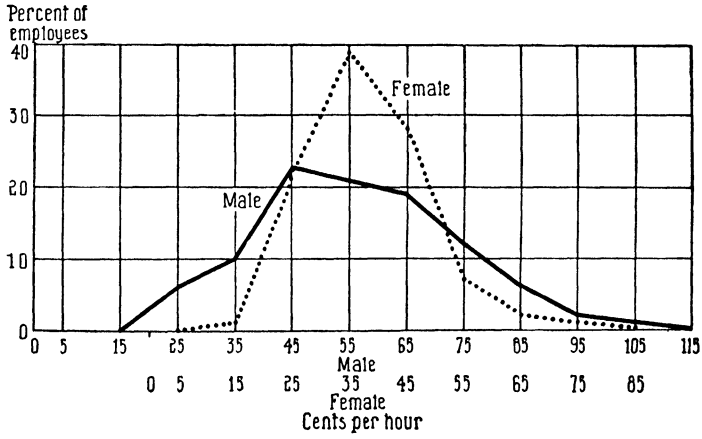


CHART 45.—Polygons of Chart 44, with "female" scale displaced laterally.

with the female scale expanded by 50 per cent. Chart 45 emphasizes the difference in spread, and Chart 46 ostensibly removes much of that difference.

The curves of Charts 44 to 46, especially the male curves, have their peaks well to the left of the center of the range of variation. This lack of symmetry can in some cases be largely removed by

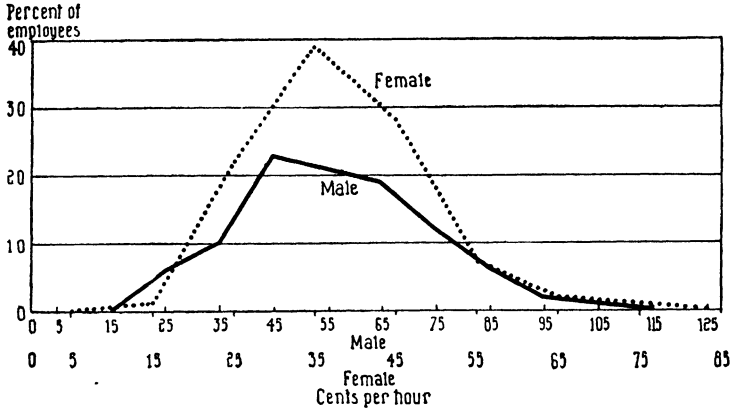


CHART 46.—Polygons of Chart 44, with "female" scale expanded.

using a horizontal ratio scale, as in Chart 47. The curve of Chart 47 is not very satisfactory for the smaller values of the variable.

TABLE 43

SYMMETRICAL CURVE CORRESPONDING TO THE "FEMALE" SERIES OF TABLE 42*

Hourly earnings[a]	Actual relative frequency (1)	Frequencies of normal curve (2)
0–10		0
10–20	1	4
20–30	22	18
30–40	39	37
40–50	28	30
50–60	7	10
60–70	2	1
70–80	1	0
Total	100	100

* See below, Chap. XIII, for more detailed treatment of column 2 and its derivation.
[a] Upper limit of class interval excluded. Unit: one cent.

This suggests that, for charting on a ratio scale, there is some advantage in using narrower class intervals for the smaller variates than are necessary for the larger. The class intervals might even

be determined so that they appear equal on a ratio scale (see below, page 226). The absence of symmetry, even for the female curve, is emphasized by Chart 48, which compares that curve with the perfectly symmetrical curve of column 2 in Table 43 (described below, Chap. XIII).

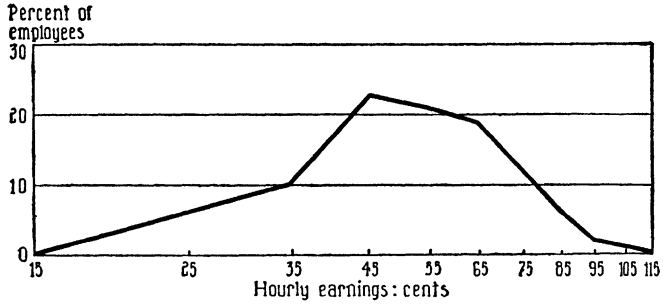


CHART 47.—Polygon of the "male" series of Chart 44, with earnings scale logarithmic.

Characteristics.—The determination of the nature and precise amounts of the several displacements and distortions applied above implies a succession of numerical computations which constitute the framework of statistical analysis for frequency series. The lateral shifting of the horizontal scale (Chart 45) implies a

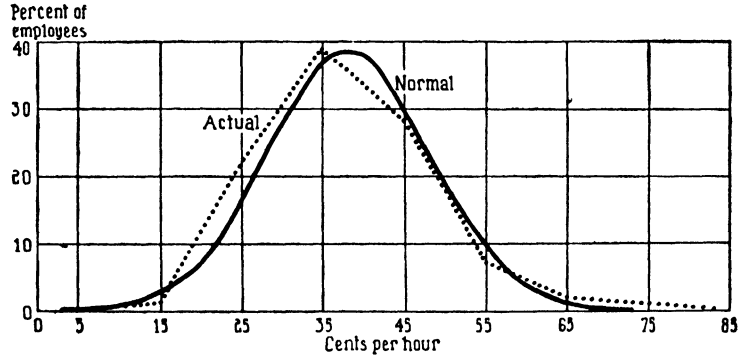


CHART 48.—Polygon of actual data, and fitted normal curve, for "female" series of Chart 44.

(Data in Table 43, page 126.)

knowledge of the *mean* of each series, the scale expansion to remove the difference in spread (Chart 46) roughly compares the *dispersion* of the two series, and the plotting on a ratio scale (Chart 47) removes some of the *skewness* in the series. The differences between the two curves in Charts 45, 46, 48 emphasize important

properties of a frequency curve. These properties are called *characteristics*, and the mean, dispersion, and skewness are, in order, the first three characteristics of a series.

The characteristics of a frequency series describe briefly the nature of the variation recorded in that series. They are summary numbers designed to present the essential features of the frequency distribution. They serve the same purpose for frequency series as summary measurements serve in describing any other complicated object or phenomenon. They effectively summarize a frequency curve just as the climate of a town is summarized by giving its mean quarterly temperature, humidity, and precipitation, and just as the agricultural industry of a community is summarized by citing the average acreage, yield per acre, and price, for each crop. The precise definitions of the several characteristics must await the discussion of numerical methods of analysis. Moreover, the above process of *graphic* adjustment is not designed to measure completely the several characteristics in succession, nor is it appropriate for examining every frequency series. In fact, the comparison of one frequency series with another, or with some standard series, by this method will often prove unsuccessful because of other and more complicated deviations from the standard than those implied by the characteristics. Furthermore, such graphic devices are almost never appropriate for presentation charts, and should be used with caution even in working charts. The process is used in the above illustrations chiefly to give an advance view of the general plan of analysis, and to suggest the place of graphic method in such analysis.

THE BLOCK DIAGRAM

The data of a frequency series can be shown effectively by a *block diagram*. Instead of plotting a point above the center of each class interval, a rectangle with height proportional to the stated frequency is erected upon each class interval as a base (see Chart 41).

If all class intervals are of equal width the areas of the several blocks (rectangles) are proportional to their altitudes. Hence, as the altitudes are proportional to the frequencies in the given series, the areas of the blocks are proportional to the frequencies. By an extension of this idea, a particular area of any shape and located anywhere on the diagram may be regarded as representing a specific frequency. In addition to affording a means of appraising certain properties of a frequency series or of making certain

comparisons between such series, this area notion establishes a basis for *smoothing* a frequency curve (below, page 133).

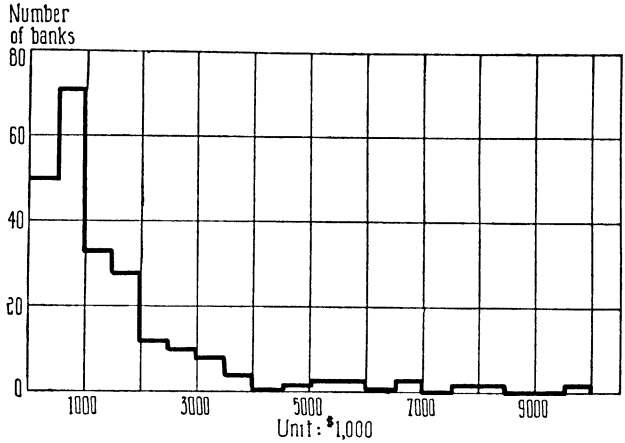


CHART 49.—Number of national banks in Ohio, excluding reserve cities, according to amount of deposits, December 31, 1936.
(Data in Table Q, Appendix A)

Chart 49 shows a block diagram for a series departing much farther from symmetry than that of Chart 41, and the correspond-

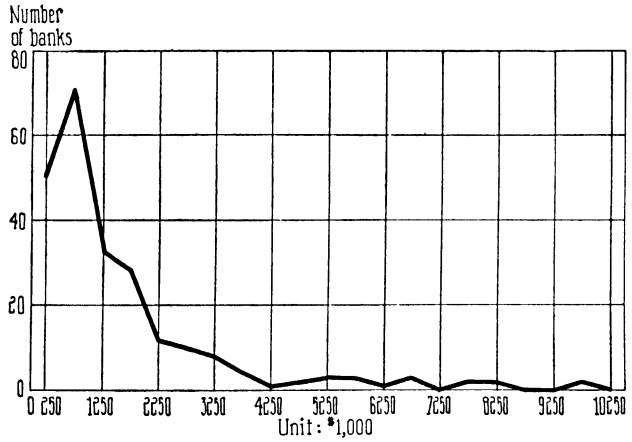


CHART 50.—Frequency polygon corresponding to block diagram of Chart 49

ing frequency polygon appears in Chart 50. Various other illustrations of block diagrams appear below, particularly in Chap. XI.

Although the plotting task is slightly easier for the polygon than for the block diagram, the latter has several advantages as an aid to study of a frequency series. In particular, as indicated above (page 122), the block diagram gives a correct representation of total frequency: the area between the horizontal axis and the block diagram is proportional to the total frequency, whereas this is not necessarily true with the polygon. (See Appendix B for suggestions which simplify the plotting of block diagrams.)

THE STATISTICAL SAMPLE, AND SMOOTHING

A frequency polygon or a block diagram accurately presents the frequency series as an isolated group of data. Often in practice, however, a given frequency series, based upon observations covering a limited group of cases, is expected to yield information concerning a larger group which includes the cases actually observed. Thus it may be desired, by a study of the statistics of the workers covered by column 1 of Table 42, to reach conclusions pertaining to all female workers in the specified occupation in the boot and shoe industry in 1930. The data actually given are then regarded as pertaining to a sample from the all-inclusive body of such workers. Again, a statistical study of the price fluctuations of a small group of commodities may serve as a basis for conclusions concerning changes in the prices of all commodities entered in trade, or at least of a more inclusive list than that covered by the actual sample. A limited group of objects, which is used in a statistical analysis to represent a larger group from which the limited group has been selected, is called a *sample* (see above, page 52). The larger group is called the *population.*

The minor peaks and dips in a frequency polygon or block diagram, while accurately exhibiting the peculiarities of the actual data, may be regarded as giving an imperfect picture of the ideal frequency curve of the larger group from which the limited group of actual data is a sample. In other words, each point on the actual curve is considered as locating approximately a point on the unknown curve belonging to the group having a larger total frequency, a group of which the observed group is a part. In this view, of course, it is advantageous to have a chart based on relative frequencies, because total frequency is less for the sample than for the population. The unknown curve for the entire population is *assumed* to be moderately regular in its form, largely on the ground that an actually observed frequency series having a high total frequency—pertaining to a large population—is generally found,

in many fields of social as well as physical and natural inquiry, to be tolerably smooth and free from irregularities. Accordingly, points (or blocks) of the actual curve, belonging to the sample, which present irregularities are taken to be relatively poor approximations—to be due to relatively large errors in the sense that they inaccurately represent the unknown curve. These inaccuracies are called *errors of sampling*, and will receive further attention below (page 218). Where the actual curve is smooth and regular, its points are assumed to be less in error, as representatives

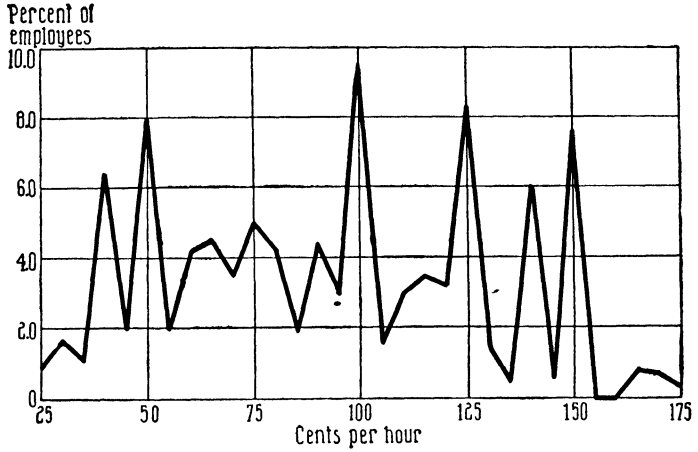


CHART 51.—Per cent of building-trades workers receiving wages within five-cent ranges having specified centers, in 1936.
(Data in Table 44, page 132.)

of points on the unknown curve, than where the actual curve turns sharply or shifts abruptly.

Both curves of Chart 44 are tolerably smooth as they stand, and we may suspect that the percentage figures are based on fairly large samples (the total frequency of which is not specified in the tabulation). The curve of Chart 40 is somewhat less smooth as it stands; and this is somewhat more apparent in the block diagram (Chart 41), the latter being a more effective means of emphasizing irregularities than the polygon. The curve of Chart 51 shows much greater irregularity, partly no doubt because of the more numerous class intervals but partly because the sample is almost surely derived from a nonhomogeneous population.[1]

[1] In Chart 51 the horizontal scale does not begin at zero, and this is an objection to the diagram. The curve is not plotted right of 175, because of an open-end ("all over") class (see Table 44).

The process of *smoothing* aims to derive, by study of the actual curve plotted from the sample data, the unknown curve belonging to the population from which the sample was presumably drawn. As in the case of time series, smoothing can be done by inspection of the actual chart or by computations implying an assumed mathematical formula for the unknown curve. Discussion of the second method must be postponed until the numerical derivation and properties of the characteristics have been studied. The first,

TABLE 44

NUMBER AND PER CENT OF BUILDING-TRADE EMPLOYEES RECEIVING EACH CLASSIFIED RATE PER HOUR, 1936*

Classified rate per hour[a]	Number[b]	Per cent	Classified rate per hour[a]	Number[b]	Per cent
All employees	186,145	100 0			
Under 22½	288	0 2	97 5–102 5	17,541	9 5
22 5–27 5	1,703	0 9	102 5–107.5	2 920	1 6
27.5–32 5	3,252	1 7	107 5–112 5	5,616	3 0
32 5–37 5	2,080	1 1	112 5–117 5	6,518	3 5
37 5–42.5	11,871	6 4	117 5–122 5	5,863	3 2
42.5–47.5	3,748	2 0	122.5–127.5	15 378	8 3
47.5–52 5	14,734	7 9	127 5–132 5	2,860	1 5
52 5–57 5	3,632	2 0	132 5–137 5	892	0 5
57.5–62 5	7,878	4 2	137 5–142 5	11,251	6 0
62 5–67.5	8,382	4 5	142.5–147 5	1,147	0 6
67.5–72 5	6,513	3 5	147.5–152.5	14,104	7 6
72 5–77 5	9,355	5 0	152 5–157.5	85	[c]
77 5–82 5	7,753	4 2	157 5–162 5	27	[c]
82 5–87.5	3,537	1 9	162 5–167 5	1,469	0 8
87 5–92 5	8,240	4 4	167 5–172 5	1,295	0 7
92.5–97.5	5,666	3 0	172 5–177 5	478	0 3
			177 5 and over	69	[c]

* Source: *Monthly Labor Review*, August, 1937, p 290.
[a] Lower limit inclusive. Unit one cent
[b] Unit: one person.
[c] Less than one-tenth of 1 per cent.

or visual, method involves only the freehand insertion of the smoothed curve after the polygon or block diagram is complete.

The area principle furnishes an excellent guide in smoothing a block diagram, and this is another reason for preferring the block diagram to the polygon. The total area between the finished smooth curve and the horizontal axis must equal the total area of the blocks, in order that the total frequency remain unchanged. Moreover, so far as possible without destroying the regularity of the curve, the curve must add to each block (the horizontally hatched quasi-triangular sections of Chart 52) approximately the same area

as it deducts (the vertically hatched sections). These ends can be attained by a careful visual estimation of the areas, but greater

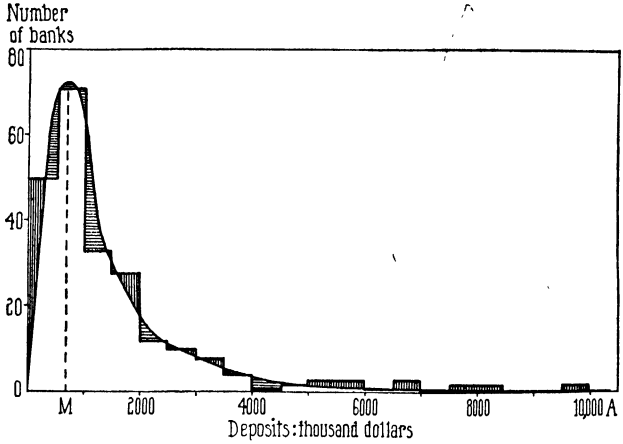


CHART 52.—Smooth curve fitted visually to the block diagram of Chart 49.

ease and accuracy are possible by the use of a finely divided grid (Chart 53). The counting of the small squares and sections of squares within each area to be estimated yields approximate

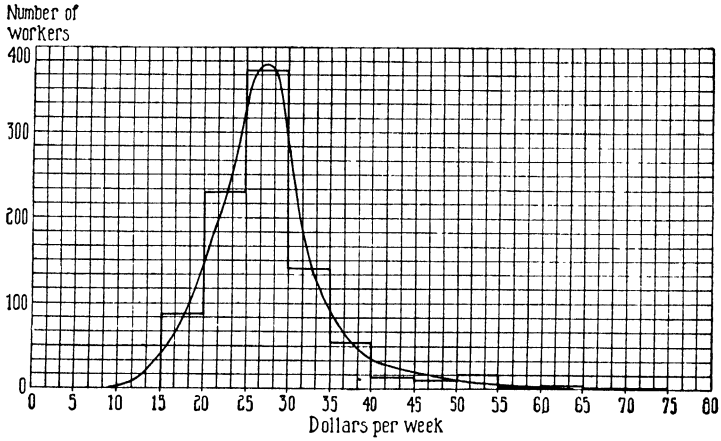


CHART 53.—Smooth curve for the block diagram of Chart 41, fitted by aid of a fine rectangular grid.

measurements which are likely to be more reliable than those made by the unaided eye.

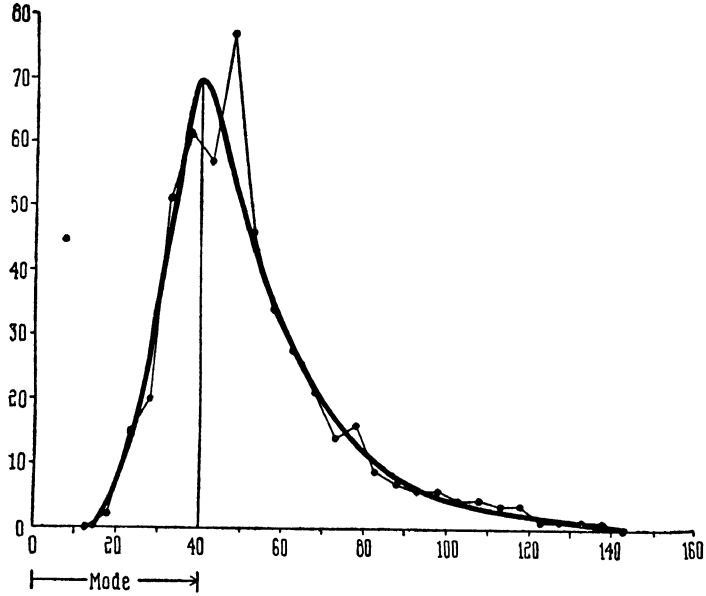


CHART 54.—Frequency polygon of number of auditors in New York City in May 1937, classified according to earnings in dollars per week; and smooth curve fitted visually.

(Data in Table 48, with adjustments.)

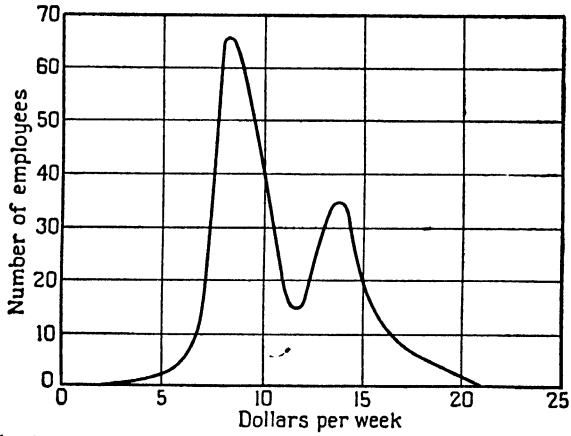


CHART 55.—Smooth curve, visually fitted to frequency distribution of wages of male machine operators in tanneries in Central States in 1900.

The frequency polygon also can be smoothed, but the result is usually less satisfactory than that obtained from the block diagram (Chart 54). In the absence of the area rule as a guide to smoothing the polygon, the estimates of form and position of the smoothed curve must rest upon considerations somewhat similar to those touching the smoothing of time graphs (Chap. VIII).

The smoothed curve—whether the smoothed polygon or block diagram—yields an estimate of some of the significant summary measures of the frequency series. The variate corresponding to the highest point on the curve is called the *mode* (*OM* in Chart 52). A curve in which there are two modes is called *bimodal;* or if there are two or more modes, *multimodal* (Charts 55 and 56). The

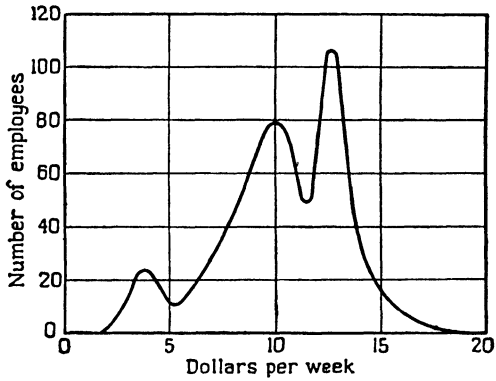


CHART 56.—Smooth curve, visually fitted to frequency distribution of wages of male machine hands, wood, in furniture mills in Central States in 1900.

difference between the lowest and highest variates is called the *range* (*OA* in Chart 52). Chart 45 suggests that the range is indicative of the amount of dispersion; but this use of the range may be misleading, for both curves may have the same range whereas one of them has the smaller dispersion as that term is defined (see Chap. XII). To put it more emphatically, although one curve has a greater range than another, the first may nevertheless have a smaller dispersion (see below, page 193).

OGIVES

An entirely different method of charting a frequency series is by the *cumulative frequency curve,* or *ogive.* Columns 3 and 4 of Table 40 give the cumulative frequencies for the series of column 2, column 3 cumulating *upward* and column 4 *downward.* Thus, each

item of column 3 gives the total frequency below the upper limit of the corresponding class interval, and each item of column 4 gives

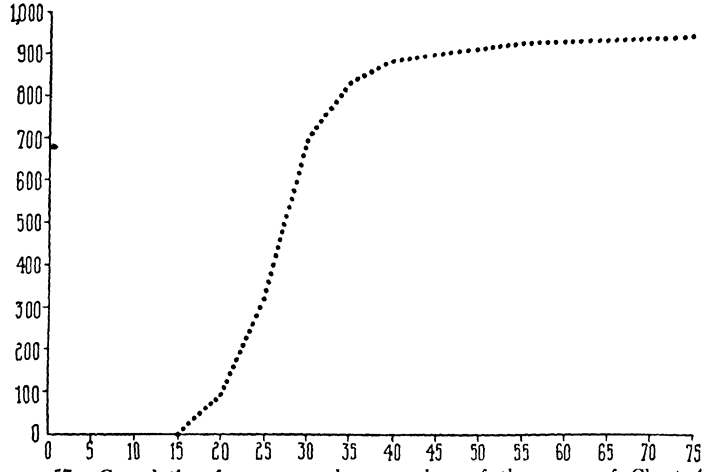


CHART 57.—Cumulative frequency polygon—ogive—of the series of Chart 40. (Units Horizontal, dollars per week; vertical, one worker. Data in Table 40, page 119.)

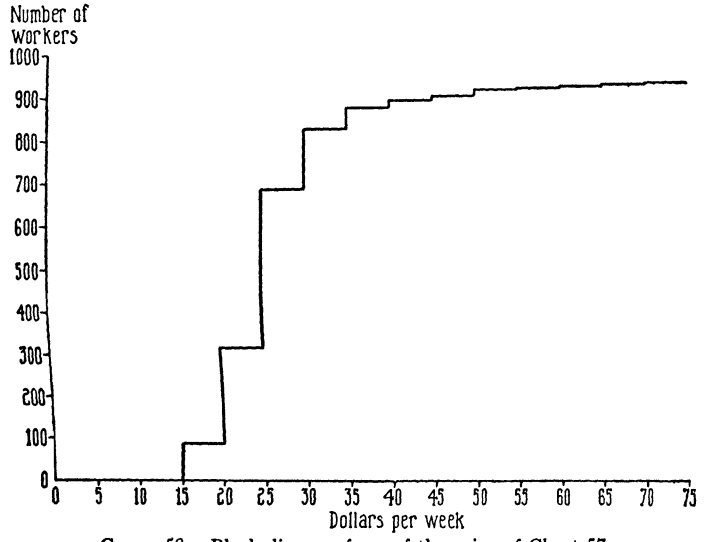


CHART 58.—Block-diagram form of the ogive of Chart 57.

the total frequency above the lower unit of the interval. Each item of column 3 *plus* the adjacent item in column 4 must equal the

total frequency, given at the foot of column 2. The ogive corresponding to column 3 appears in the form of a polygon in Chart 57, and in block form in Chart 58. The ogive (polygon type) for column 4 appears in Chart 59.

Both ogives are shown in Chart 60. The highest point on either ogive gives the total frequency of the series. This point can be projected upon the vertical axis, as indicated by the dashed line near the top of the chart; and the *intercept* (distance from zero to the point of intersection) on that axis then measures the total

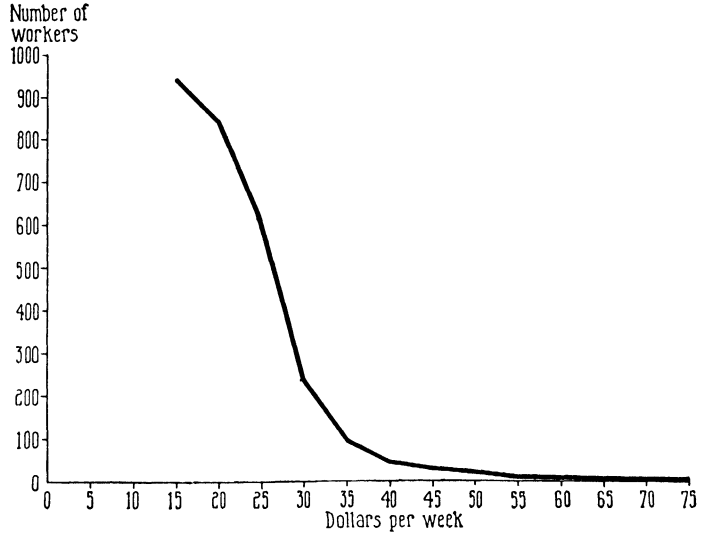


CHART 59.—Ogive, for the basic series of Chart 57, with cumulation in inverse direction.

(Data in Table 40, page 119.)

frequency. This intercept can be divided into halves, quarters, tenths, or hundredths, or any other desired portions, and the points of division projected horizontally to the right upon either ogive and then vertically downward upon the horizontal axis. Such projections are carried out for selected fractions of the total frequency, and indicated by dotted lines and arrows in Chart 61. The distance from 0 to the point where one of the vertical projections intersects the horizontal axis—the "intercept" corresponding to that point of intersection—is a hypothetical value of the variable having a specific meaning. For example, the intercept $0C$ in Chart 61, obtained by selecting A so that $0A$ is one-fourth of the

total frequency, projecting horizontally to the ogive at *B* and then vertically to the axis at *C*, is called the *lower quartile* of the series. The lower quartile is the variate below which one-fourth of the frequency falls. The variates corresponding to the various other points of intersection of the projection lines with the horizontal axis also have particular meanings. Thus *E*, corresponding to *D* which cuts off half the total frequency, is called the *median*. *G*, corresponding to *F* which cuts off three-quarters of the total

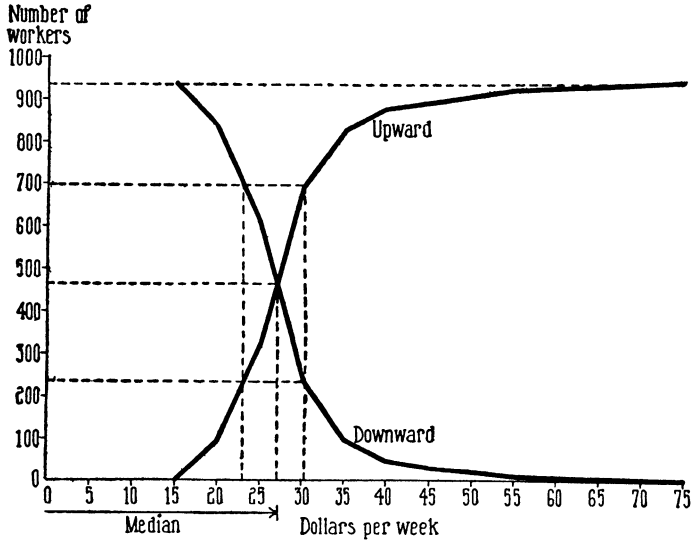


CHART 60.—Both ogives, of Charts 57 and 59, with indicated median. (Data in Table 40, page 119.)

frequency, is called the *upper quartile*. *K*, corresponding to *H* which cuts off seven-tenths of the total frequency, is called the seventh *decile*. There are, of course, nine different deciles; and they divide the total frequency into ten equal parts. Similarly, *percentiles* are values of the variable—variates—obtained by dividing the total frequency into hundredths. Further discussion of the median, quartiles, and other similar magnitudes will be presented in connection with their arithmetical determination (Chaps. X to XII).

The graphic determination of these magnitudes should be made from a smoothed ogive, rather than from the polygon as in Chart 61. But, where the polygon is as regular as in Chart 61, the results would be only slightly different. The process of smoothing

an ogive—polygon or block form—is analogous to that used in smoothing an ordinary frequency diagram.

The projection process of the graphic determination can, of course, be carried out for either ogive, although Chart 61 shows it only for the upward cumulation case. Where the two ogives are smooth—or nearly smooth, as in Chart 60—the results obtained from the two ogives are substantially equal. For ogives not smooth or nearly smooth, the two sets of results may differ sub-

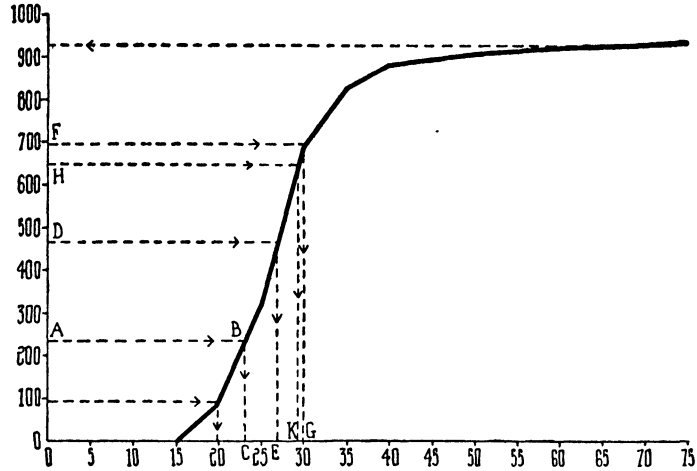


CHART 61.—Graphic location of median, and certain other summary numbers, for the ogive of Chart 57.

stantially. It should be noted also that the point of intersection of the upward and downward ogives (Chart 60) gives approximately the median of the series. The point of steepest incline on either smooth ogive approximately locates the mode (Chart 61).

DISCRETE FREQUENCY SERIES

Thus far the frequency series have been *grouped* frequency series, in which the variable is divided into class intervals according to which the frequencies are grouped. A frequency series can also arise in which each of the several items gives the frequency, not for an *interval* but for an *isolated value*, of the variable—as in the "number of issues" columns in Table 45. Such a series is *discrete* in form, whether because of a discontinuous process of measurement, perhaps resulting from custom (Table 46), or of a really discontinuous variable (Table 45). In arithmetical analysis the

TABLE 45

SAMPLE DATA ON AVERAGE COUPON RATES CARRIED BY INDUSTRIAL LONG-TERM OBLIGATIONS IN 1922*

Coupon rate %	Manufacturing			Mining and quarrying			Other industrial		
	Funded debt	Computed interest charge	Number of issues	Funded debt	Computed interest charge	Number of issues	Funded debt	Computed interest charge	Number of issues
3½	2,047	72	1						
4	8,710	348	7	71,755	2,870	6	2,500	100	1
4½	38,545	1,735	6	11,551	520	4	1,450	65	1
5	1,043,390	52,370	160	250,354	12,518	70	25,805	1,290	32
5½	140,596	7,733	23	3,215	177	3	35,169	1,934	7
6	729,739	43,784	457	343,356	20,601	136	159,943	9,597	108
6½	160,420	10,427	67	16,570	1,077	5	15,115	982	12
7	644,786	45,135	298	166,747	11,672	60	80,114	5,608	37
7½	245,924	18,444	72	19,220	1,442	7	7,440	558	6
8	320,066	25,605	187	52,001	4,160	35	34,043	2,723	16
10	.....	.....	.....	1,000	100	1			
Total	3,338,223	205,653	1,278	935,769	55,137	327	361,579	22,857	220
Average coupon rate	.....	6 16	.....	.....	5 89	.....	.....	6 32	.....

* Unit for value figures, thousand dollars. Source: "Long-Term Debts in the United States," Washington, U. S. Department of Commerce, 1937, p 95

TABLE 46

NUMBER OF NATIONAL BANKS IN MICHIGAN HAVING STATED CAPITAL ON DECEMBER 31, 1936*

Capital[a]	Frequency	Capital[a]	Frequency
25	2	250	1
35	2	275	1
37 5	2	290	1
45	1	300	1
50	24	350	1
62 5	1	400	2
75	1	460	1
80	1	490	1
100	17	500	1
110	1	540	1
122	1	550	1
142 5	1	600	1
150	2	650	1
164 8	1	910	1
165	2	988	1
200	1	1000	1
210	1	1250	1
		Total	79

* Compiled from Table 61.
[a] Unit: thousand dollars.

distinction between continuous (or grouped) and discontinuous (or discrete) series is of importance, and the consequences are discussed below (Chaps. X to XII). Sometimes, as in Table 47, a series is classified by both intervals and isolated values; but, as in this case, such a combination of discrete and continuous classifications usually reflects the fact that custom leads to a concentration of most of the variates at particular values. Analysis of such hybrid series is so difficult that they are of little use except in tabular exhibits.

TABLE 47

DISTRIBUTION OF BUILDING-TRADE WORKERS CLASSIFIED ACCORDING TO FULL-TIME WEEKLY HOURS, 1936*

Weekly hours[a]	United States[b]	Middle[b] Atlantic	East North[b] Central	South[b] Atlantic	Pacific[b]
30	17,221	5,244	2,791	780	4,464
Over 30 and under 36	2,655	470	348	119	263
36 and under 40	298	....	12	8	102
40	131,588	36,903	31,801	16,950	13,665
Over 40 and under 44	86	10	35	15	2
44	19,185	2,295	3,433	5,198	1,010
Over 44 and under 48	1,010	56	369	303	24
48	9,196	258	1,373	474	2,156
Over 48 and under 54	2,550	52	598	1,434	. ...
54 and under 60	1,582	228	373	359	95
60 and over	774	390	47	38	.....

* Source: *Monthly Labor Review*, October, 1937, p. 792.
[a] Unit: one hour. [b] Unit: one worker.

In charting, also, the distinction between discrete and continuous series has practical significance. A block diagram is scarcely a fitting diagram of discrete data, although separate bars are sometimes used (Chart 62). Moreover, the line segments in the polygon should not be used to "interpolate" between the plotted points when the observed variable is truly discrete (Chart 63), and can be used only after careful consideration when the series is discrete merely because of limitations in measuring the variable (Chart 64). In fact, the construction of a frequency polygon for a discrete series can seldom be justified except as a means to assist the eye in following the plotted points, and for that reason the plotted points should bear spots and the lines should be drawn light. In the continuous series, on the other hand, interpolation is possible for any well-chosen class interval (Chart 65).

The exact location of the left-hand and right-hand zero frequencies may be uncertain in the polygon of a discrete series

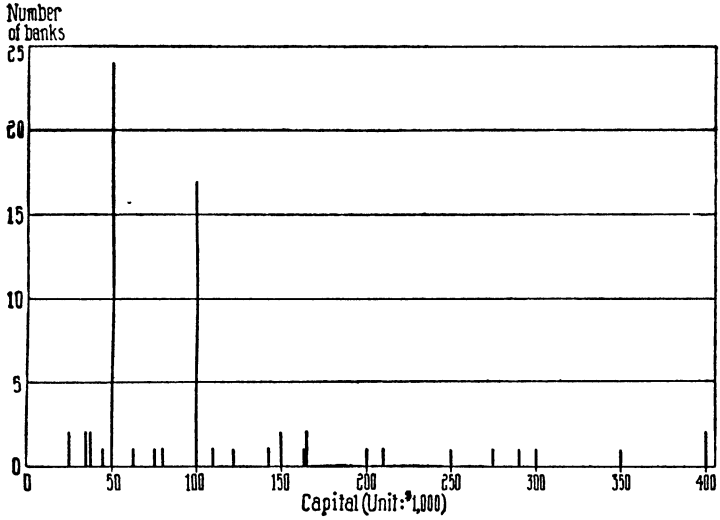


CHART 62.—Number of national banks in Michigan having stated capital on December 31, 1936.

(Data in Table 46, page 140.)

having unequally spaced values—whether 4 or 4½, or 8½ or 9, or some other value at each end (in Chart 66). Unless the specified

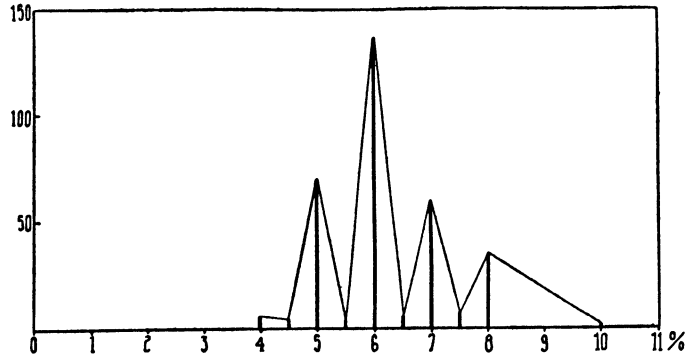


CHART 63.—Number of long-term obligations of mining and quarrying companies bearing stated rates of interest, in 1922.

(Data in Table 45, page 140.)

values of the variable are equally spaced in a discrete series, the frequency polygon may give especially misleading impres-

sions if any attention is given to a rough appraisal of the areas under the broken line (Chart 66). Here the area under the left

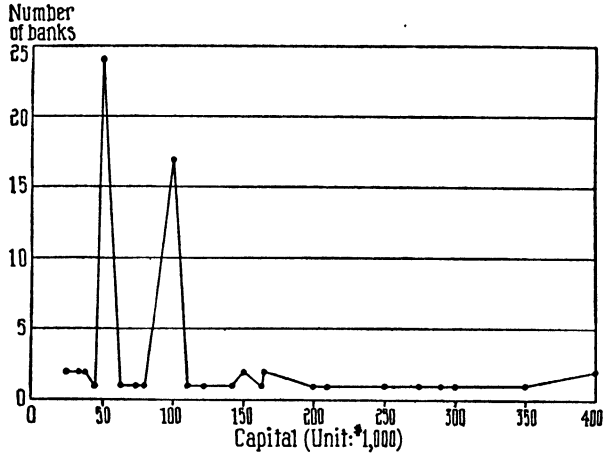


CHART 64.—Discrete series of Chart 62, with connecting line segments.

half of the polygon is not much different from what it would be if there were *no* bonds at 5½. A somewhat similar limitation holds

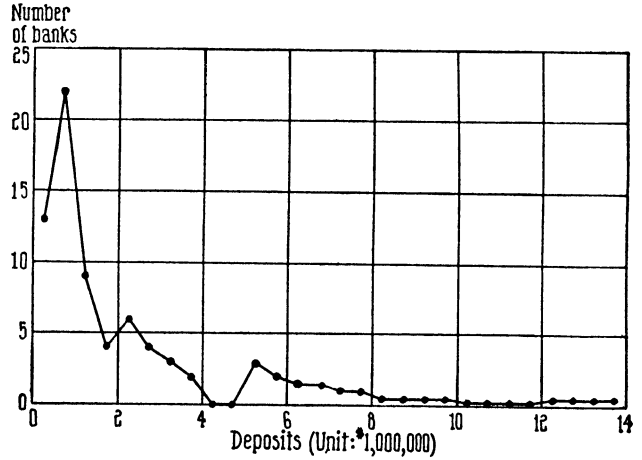


CHART 65.—Distribution of Michigan national banks according to amount of deposits on December 31, 1936, with line segments as basis of interpolation. (Data in Table *R*, Appendix A.)

for the frequency polygon of a continuous series in which the intervals vary in width (Chart 67), but an adjustment of the

plotted frequencies can readily be made to correct this false impression (Chart 68, wherein the extreme variate has been omitted for convenience). The nature and basis of this adjustment, as carried out in the right column of Table 48, are more evident when considered from the area point of view (Chart 69).

TABLE 48

DISTRIBUTION OF AUDITORS, IN A SAMPLE OF NEW YORK CITY CLERICAL WORKERS, IN MAY, 1937, ACCORDING TO WEEKLY EARNINGS*

Weekly earnings[a]	Number	Width of interval[b]	Equivalent frequency[c]
15– 20	2		
20– 25	15		
25– 30	20		
30– 35	51		
35– 40	61		
40– 45	57		
45– 50	77		
50– 55	46		
55– 60	34		
60– 65	28		
65– 70	21		
70– 75	14		
75– 80	16		
80– 85	9		
85– 90	7		
90–100	12	2	6
100–110	9	2	4 5
110–120	7	2	3 5
120–130	2	2	1
130–140	2	2	1
.....	...		..
175–200	0	5	0
200–225	1	5	2
225–300	0	15	0
Total	491		

* Source: *Monthly Labor Review*, January, 1938, p. 215.
[a] Upper limit of class interval excluded. Unit: one dollar.
[b] Width of broader intervals in terms of the 5-dollar interval.
[c] Average frequency per 5-dollar interval, in broader intervals.

WEIGHTED FREQUENCIES

Strictly, the frequency in a frequency series represents the *number* of cases having the variable of a specified size or within a specified interval. By an extension of the notion of frequency, however, a corresponding series can be obtained in which the statistical item is no longer a mere number of cases but that

number *weighted* by some specified factor. Thus, whereas the items of the middle column of Table 49 give the mere number of

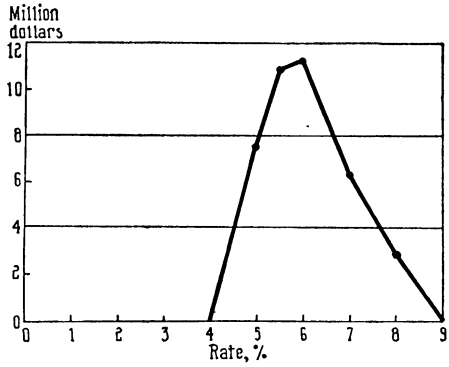


CHART 66.—Funded debt of the Republic Steel Corporation, in hands of public on December 31, 1933.
(Compiled from 1933 *Annual Report* of Company, page 11.)

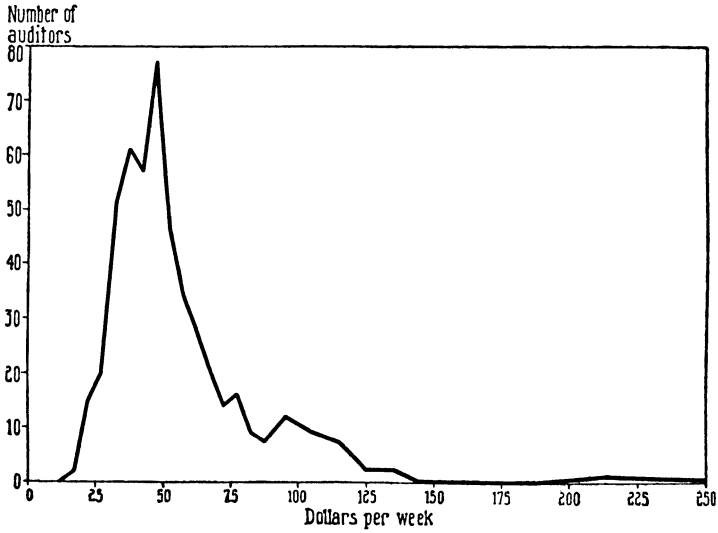


CHART 67.—Number of auditors having stated wages in New York City in May 1937; frequency polygon without adjustment for unequal class intervals.
(Data in Table 48, page 144.)

commodities—frequency in the strict sense—having price relatives within the stated intervals, the items of the right column give the total weight assigned to all the commodities within a specified

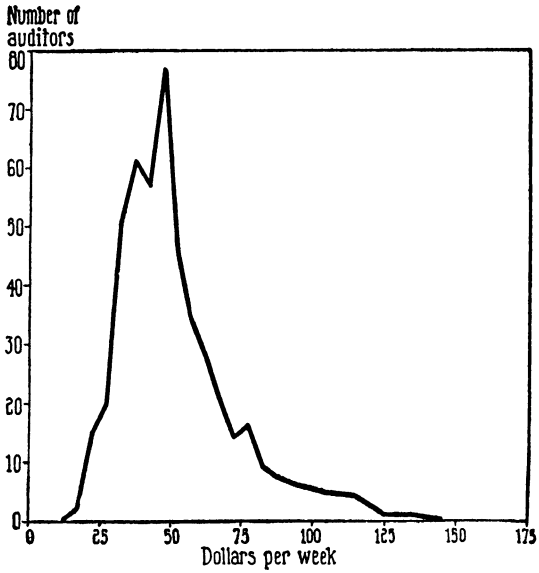


CHART 68.—Polygon of Chart 67—except for extreme right portion—after adjustment for unequal class intervals.

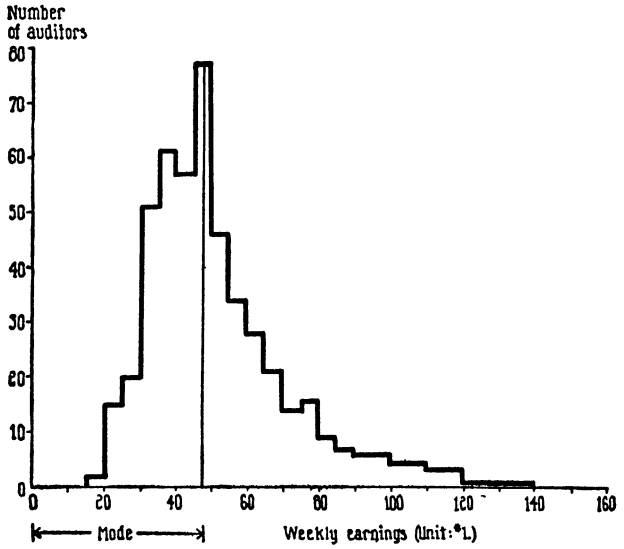


CHART 69.—Block diagram corresponding to adjusted polygon of Chart 68.

interval (see below for discussion of the weighting of variates, page 163, and of price index numbers, page 283). Chart 70 shows the two corresponding curves—the solid curve for frequencies in the strict sense, the dotted curve for weighted frequencies. Sometimes

TABLE 49

FREQUENCY DISTRIBUTION IN 1937 OF THE INDEXES OF INDIVIDUAL COMMODITIES INCLUDED IN THE "FOODS" GROUP OF THE U. S. BUREAU OF LABOR STATISTICS PRICE INDEX, ACCORDING TO THE ESTIMATED VALUES OF THE COMMODITIES IN 1926*

Price relative[a]	Number of commodities	Estimated value[b]
20 0	1	8
. . . .	. .	. . . .
40 0	1	10
45 0	4	331
50 0	1	16
55.0	2	60
60 0	8	481
65 0	3	134
70 0	6	405
75 0	40	2955
80 0	11	2148
85 0	12	1143
90 0	8	228
95.0	7	646
100 0	3	174
105 0	2	212
110 0	3	287
115 0	4	180
120 0	0	0
125 0	2	400
130 0	1	790
. . . .	. .	. .
150 0	2	21

* Based upon the indexes for the year 1937 given in the mimeographed report *Average Wholesale Prices and Index Numbers of Individual Commodities, December and Year* 1937, U. S. Bureau of Labor Statistics, and the 1926 estimated values of individual commodities in "Wholesale Prices in 1931," Washington, U. S. Bureau of Labor Statistics, Bulletin 572, January, 1933, pp. 93–96.

[a] Lower limit of 5 per cent interval is stated. Unit: per cent of 1926 average.

[b] Unit: million dollars. Estimated value in exchange for 1926.

weighting tends to smooth a frequency curve, and sometimes the reverse. Chart 70 illustrates both tendencies: in the range 70 to 90, the dotted curve is somewhat less irregular than the solid curve; but just to the right of 90 and at the right end of the chart, the reverse is true. The danger that the weighted curve will be less smooth is always present where a particular case—a particular

commodity, in the illustration—may have a weight much above, or much below, the general average weight.

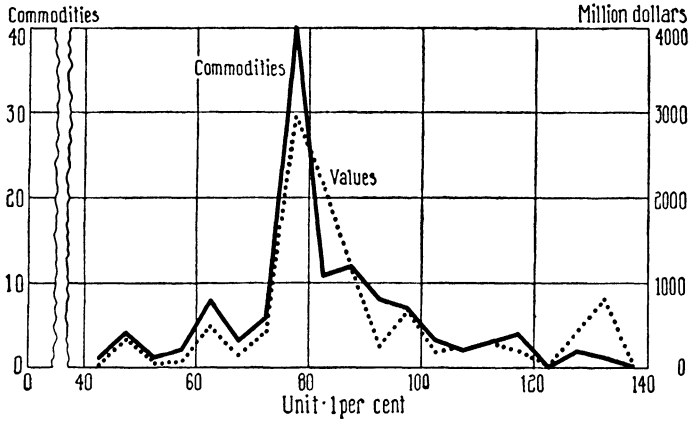


CHART 70 —Frequency polygons, according to number of commodities and values in exchange, for price relatives 1937/1926 of commodities of the foods group of the U. S. Bureau of Labor Statistics price index.
(Data in Table 49, page 147.)

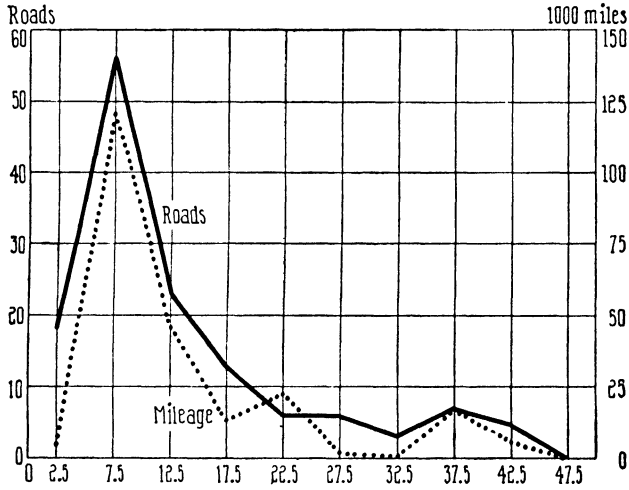


CHART 71.—Frequency polygons, according to number of roads and mileage, of Class I railroads classified by freight revenue per mile in 1936.
(Unit: thousand dollars per mile. Data in Table 50A, page 149.)

Further illustrations of weighted frequencies, compared with ordinary frequencies, appear in Charts 71 and 72. In these cases,

the weights—miles of road for each carrier—are the same as the denominator of the variate, which is revenue per mile of road

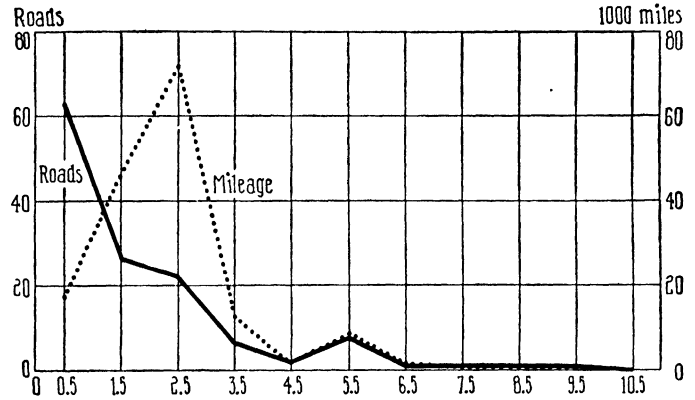


CHART 72.—Frequency polygons, according to number of roads and mileage, of Class I railroads classified by passenger revenue per mile in 1936.
(Unit: thousand dollars per mile. Data in Table 50B, page 150.)

(Table 50). Only indirectly and to a limited extent—in that both the weight, aggregate exchange value at a specified time, and the

TABLE 50A
FREQUENCY DISTRIBUTION OF FREIGHT REVENUE PER MILE FOR CLASS I RAILWAYS OF THE UNITED STATES IN 1936*

Revenue[a]	Number of roads	Number of miles
0 0- 4.9	18	4,856
5.0- 9.9	56	121,228
10.0- 14 9	23	46,642
15 0- 19.9	13	13,440
20.0- 24.9	6	22,549
25.0- 29.9	6	1,885
30 0- 34 9	3	1,108
35.0- 39.9	7	16,778
40.0- 44.9	5	6,030
.....	.	.....
70.0- 74.9	1	217
75.0- 79.9	1	50
.....	..	.....
90.0- 94.9	1	234
.....	..	.....
135.0-139.9	1	20

* Source: "Statistics of Railways in the United States, 1936," Washington, Interstate Commerce Commission, 1937, pp. 16–144.
[a] Unit: thousand dollars per mile of road.

denominator of the variate, price relative (stating price in one year as a percentage of the price in another), contain a price factor—was this true of Table 49. In Chart 71 both curves give the same general picture of the distribution, although the weighted curve is the less smooth. In Chart 72 a much more striking difference

TABLE 50B

FREQUENCY DISTRIBUTION OF PASSENGER REVENUE PER MILE FOR CLASS I RAILWAYS OF THE UNITED STATES IN 1936*

Revenue[a]	Number of roads	Number of miles
0.0– 0.9	63	17,136
1.0– 1.9	26	46,179
2.0– 2.9	22	71,797
3.0– 3.9	6	12,317
4.0– 4.9	2	1,048
5.0– 5.9	8	8,873
6.0– 6.9	1	1,640
7.0– 7.9	1	684
8.0– 8.9	1	396
9.0– 9.9	1	343
10.0–10.9	0	0
11.0–11.9	2	9,125
12.0–12.9	1	46
13.0–13.9	2	7,470
.....	..	.....
28.0–28.9	1	1,147
29.0–29.9	1	117
.....	..	.....
32.0–32.9	1	5
.....	..	.....
42.0–42.9	1	23
.....	..	.....
52.0–52.9	1	360

* Source: "Statistics of Railways in the United States, 1936," Washington, Interstate Commerce Commission, 1937, pp. 16–144.
[a] Unit: thousand dollars per mile of road.

appears: the weighted curve has a well-defined mode, whereas the unweighted curve is J-shaped.

The weighting factor can also be related directly to the variate —the weight may be merely the size of the variable itself. Thus, for Table 71 the strict frequency is the number of tax returns within a specified interval. The weighted frequency is the total *amount* of dividends reported on those returns. This amount can be regarded as the product of the number of returns within the interval by the average dividends, per return, received in that

interval. As the average dividends per return must lie somewhere within the particular interval, and presumably not far from the center (class mark), the weight in this case is clearly the variate itself. If the series were discrete, this would be exactly true; and it is a close approximation to truth even for a grouped series such as that of Table 71.

The analysis of a weighted frequency series—particularly where the weights are related to the variates, directly as in Table 71 or through the denominator as in Table 50—is often of great usefulness. The analyst must, however, guard against carrying over inferences from the weighted series to the unweighted series. The striking changes which weighting can cause, as illustrated in Chart 72, warn against applying conclusions drawn from the weighted series to the unweighted series.

PART II

GENERAL ANALYTICAL METHODS

CHAPTER X

SUMMARY NUMBERS

GENERAL PROPERTIES OF AVERAGES

Part I dealt with the source and nature of statistical data, with the organization and presentation of such data, and with forms of summarization and interpretation which involve only tabulation and charting of the actual data or of items promptly obtainable therefrom. It is the object of Part II to outline those *analytical* methods of summarization and interpretation which are most commonly in use, not merely describing those methods, but explaining the assumptions upon which their use must rest and setting forth the limitations to their practical application. The present chapter is designed to give a preliminary view of some of the basic arithmetical methods; and, although the treatment covers chiefly the analysis of categorical series, the essential ideas involved are generally applicable and can with few exceptions be extended to series of other types.

One of the commonest summary numbers for a series is the *average*. The average of a statistical series is a single value of the variable, which is a satisfactory representative, for the purpose in hand, of all values of the variable (variates) included in the series.[1] In other words, an average is a *typical value* of the variable, and that particular value which is most typical for the use to which it

---

[1] The following definition is given in the eleventh edition of the *Encyclopaedia Britannica*, under the heading "Probability": "An average may be defined as a quantity derived from a given set of quantities by a process such that, if the constituents become all equal, the average will coincide with the constituents, and the constituents not being equal, the average is greater than the least and less than the greatest of the constituents. For example, if $x_1, x_2, \ldots, x_n$ are the constituents, the following expressions form averages (called respectively the *arithmetic, geometric,* and *harmonic* means):

$$\frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$(x_1 \times x_2 \times \cdots \times x_n)^{\frac{1}{n}}$$

$$\frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right)}$$

The conditions of an average are likewise satisfied by innumerable other symmetrical

will be put. The average of a particular series may be actual or artificial: it may be a real variate actually specified in the given series or it may be an intermediate value of the variable which does not actually occur in the data. For example, one average of the series in Table 51 (the median), 3,060, is an actual value of the variable; whereas another average (the mean), 4,318, is an artificial value in that none of the actual variates has that value. The *median*, in the case of a categorical series having an odd number of items, is the middle item when the items have been arranged in the order of their size. The *mean* is the result of dividing the sum of all the items by their number. That the use to which it is put determines which average is most typical appears from consideration of Table 52. In the left column of this table the mode group, 20–24, differs from the median group, 30–34, because of the numerous cases scattered through the high age groups. The *mode—apparent* mode (see below, page 186)—occurs in the interval, for a frequency series having equally wide intervals, with the maximum frequency. The *median*, for a frequency series, is defined as above but must be determined by a special technique (see Chap. XI). If one is interested in knowing the most probable age of an unemployed individual chosen at random from the 253,334 cases under study in the table, the mode should be used; but to indicate the limit of age for about half the entire group under study, the median is needed.

Because of the great variety of possible averages, defined in various ways, but all subject to the above general definition, some basis of selection of the particular type of average suitable for a

---

functions, for example:

$$\left(\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}\right)^{1/2}$$

The conception may be extended from symmetrical to unsymmetrical functions by supposing any one or more of the constituents in the former to be repeated several times Thus, if in the first of the averages above mentioned (the arithmetic mean) the constituent x_r occurs p times, the expression is to be modified by putting px_r for x_r in the numerator, and in the denominator, for n, $n + p - 1$. The definition of an average covers a still wider field. The process employed need not be a *function* [that is, an algebraic function]. One of the important averages is formed by arraying the constituents in the order of their magnitude and taking for the average a value which has as many constituents above it as below it, the median. The designation is also extended to that value about which the greatest number of constituents cluster most closely, the 'center of greatest density,' or (with reference to the geometrical representation of the groupings of the constituents) the greatest ordinate, or, as recurring most frequently, the mode. But to comply with the definition there must be the added condition that the mode does not occur at either extremity of the range between the greatest and the least of the constituents."

given series is necessary. Yule and Kendall propose six requisites of a good average, and the following discussion is based largely upon their more elaborate exposition.[1]

(1) The average should be *rigidly defined:* not an arbitrary estimate by an individual observer, but an unmistakable property

TABLE 51

TAXES AND SPECIAL ASSESSMENTS OF CLASS I RAILROADS, BY STATES, IN 1935*

Maine, 43	1,232	Missouri, 24	3,336
New Hampshire, 44	718	North Dakota, 38	2,114
Vermont, 47	413	South Dakota, 37	2,191
Massachusetts, 17	3,964	*Nebraska, 25 (median)*	*3,060*
Rhode Island, 46	664	Kansas, 8	6,232
Connecticut, 45	711	Kentucky, 22	3,552
New York, 1	23,685	Tennessee, 29	2,558
New Jersey, 2	17,520	Alabama, 34	2,378
Pennsylvania, 5	9,342	Mississippi, 28	2,937
Delaware, 48	151	Louisiana, 16	4,087
District of Columbia, 49	138	Texas, 10	5,487
Maryland, 41	1,748	Oklahoma, 21	3,667
Virginia, 12	5,192	Arkansas, 35	2,290
West Virginia, 6	8,229	Montana, 14	4,460
North Carolina, 20	3,696	Wyoming, 42	1,671
South Carolina, 33	2,396	Colorado, 23	3,363
Georgia, 31	2,434	New Mexico, 39	1,883
Florida, 27	3,039	Arizona, 26	3,055
Ohio, 4	10,452	Utah, 36	2,243
Indiana, 7	6,520	Nevada, 40	1,756
Illinois, 3	13,610	Idaho, 30	2,516
Michigan, 11	5,346	Washington, 18	3,882
Wisconsin, 13	4,921	Oregon, 32	2,413
Minnesota, 15	4,291	California, 9	6,229
Iowa, 19	3,804		

* Unit thousand dollars. Source "Statistical Abstract of the United States, 1937," Washington, U S Department of Commerce, 1938, p. 385. Numeral indicates order, ranked by amount.

of the series. Although an estimate may at times be more appropriate than a rigidly calculated average, to typify the series and facilitate comparisons with other series a strict mathematical computation is needed. Certainly the average should not be a quantity which varies with each computer or computation. (2) The average should *depend upon every item* in the series. Otherwise the use of the average to typify the whole series may rightfully be questioned. (3) The average should be simple and *easy to*

[1] G. U. YULE and M. G. KENDALL, "An Introduction to the Theory of Statistics," Chas. Griffin & Co., Ltd., London, 1937, p. 113.

understand. There are averages which, though difficult to define in terms familiar to the layman, are quite important in specialized lines of work. An example of this is the exponential average (see below, page 192), which is used quite generally in the computation of equivalent ages, and in Lidstone's Z-method for evaluating reserves under a group of insurance policies. If, however, the average is to be used very widely for presenting results

TABLE 52

NUMBER OF PERSONS UNEMPLOYED, BUT ABLE TO WORK AND LOOKING FOR A JOB, IN THE UNITED STATES, CLASSIFIED ACCORDING TO AGE, SEX, AND PERIOD OF IDLENESS, 1930*

Age[a]	Male[b]		Female[b]	
	1–2 weeks	14–26 weeks	1–2 weeks	14–26 weeks
10–19	27,284	50,334	12,277	15,331
20–24	44,910	86,268	13,520	15,393
25–29	33,929	61,645	8,216	9,032
30–34	27,003	53,327	5,570	6,507
35–39	26,560	57,020	5,211	6,297
40–44	23,711	55,096	3,912	5,117
45–49	21,980	53,737	3,400	4,259
50–54	17,200	46,663	2,595	3,257
55–59	12,821	37,547	1,794	2,305
60–64	9,070	28,495	1,103	1,509
65–69	5,334	18,342	642	816
70 and over	3,349	10,837	285	395
Unknown	183	363	34	42
Total	253,334	559,674	58,559	70,260

* Source: "Fifteenth Census of the United States, 1930, Unemployment, Vol. II," Washington, U. S Department of Commerce, 1932, p. 329.
[a] Unit: one year.
[b] Unit: one person.

to the lay public, it should not be perplexing in definition or abstruse in significance. (4) The average should be *easy to calculate*, in order to avoid waste of time and money in the computation of a needlessly complicated and difficult average. (5) The average should be relatively *free from fluctuations due to sampling.* The concept of statistical *sample* was introduced above (page 51); but a discussion of the nature of fluctuations due to sampling must be postponed to Chap. XIII, where the significance of this fifth requisite of an average will become apparent. (6) The average should *lend itself readily to further algebraic treatment.* This last property is of little importance in the elementary arith-

metical summarization of simple series, but, in the more or less exhaustive analyses of frequency data, it often becomes the controlling consideration in choosing the type of average.

TYPES OF AVERAGES

In general the arithmetic average surpasses other types in fulfilling the above requirements. So great indeed are its relative advantages that it is usually chosen unless specific evidence exists favoring the use of some other type. The *arithmetic average* of a statistical series is the quotient obtained by dividing the sum of all the variates in the given series by the number of such variates. This type of average is generally called the *mean*, and although that term is often used generically to denote any sort of average, its use in this text will be confined to this specific type. The computation of the mean, by direct use of the definition, was carried out above for the data of Table 51: the mean of the series is therefore 4,318. Similarly, for the data of Table 53, and those of Table 54, direct calculation of the mean by summing the items and dividing by their number has been carried out. Clearly, the mean can be found from a mere knowledge of the number of objects and the total of the different variates, even if the individual variates are not known. Thus the total pay roll of the United States Steel Corporation in 1936 was $338,866,121, and the number of employees was 222,372 (Corporation's *Annual Report* for 1936, page 5). Hence the mean was $1,525. A mean thus calculated, from a mere knowledge of an aggregate and of the number of items covered, is not, however, dependably typical of the various items. Without knowledge as to the frequency distribution of the items, the typicalness of the mean—or any other average—remains in doubt (see below, page 162).

The *median* (see page 138 for a graphic definition) is a value of the variable such that as many actual values lie above as below it. The median for Table 53 is 84. In all but the simplest cases, the determination of the median involves arranging the series in an *array*. A series is said to be arrayed if the variates (values of the variable) are arranged in the order of their magnitude. Thus Table 55 shows the data of Table 53 in an array. The median of an array can be found by inspection; it is the middle variate if the number of cases is odd (Table 55), and the mean of the two middle values if the number of cases is even (Table 54). Obviously, the median is an actual variate in the first case; but in the second case it is an artificial value of the variable, no given variate being the

TABLE 53

ADJUSTED INDEXES OF MANUFACTURING PRODUCTION, BY INDUSTRIES DECEMBER, 1937*

Serial number	Industrial class	Relative	Serial number	Industrial class	Relative
1	Pig iron	50	20	Goat and kid leathers[d]	74
2	Steel ingots	49	21	Boots and shoes	93
3	Cotton consumption	88	22	Cement	71
4	Wool consumption[a]	54	23	Glass plate	108
5	Wool[ab]	66	24	Tin deliveries	88
6	Carpets and rugs[ab]	43	25	Zinc	108
7	Silk deliveries	69	26	Lead	87
8	Hogs slaughtered[c]	72	27	Gasoline[e]	255
9	Cattle slaughtered[c]	101	28	Kerosene[e]	108
10	Calves slaughtered[c]	114	29	Fuel oil[e]	140
11	Sheep slaughtered[c]	143	30	Lubricating oil[e]	123
12	Wheat flour	88	31	Coke, by-product	89
13	Sugar meltings	142	32	Coke, beehive	11
14	Newsprint production	63	33	Tires, pneumatic	70
15	Newsprint consumption	134	34	Inner tubes	42
16	Automobiles	78	35	Cigars	74
17	Locomotives	16	36	Cigarettes	247
18	Cattle hide leathers[d]	76	37	Manufactured tobacco	84
19	Calf and kip leathers[d]	73			

Sum = 3395; Mean = 92

* Unit per cent of 1923–1925 average, adjusted for seasonal variation except for items 5, 24, 27, 29, 30, 33, 34 Source *Federal Reserve Bulletin*, April, 1938, p. 315.

[a] Listed under wool group
[b] Machine activity [c] Listed under slaughtering and meat packing.
[d] Listed under tanning [e] Listed under petroleum refining.

TABLE 54

INDEXES OF FACTORY EMPLOYMENT IN FOOD-PRODUCTS INDUSTRIES, 1937*

1.	Beverages	204 9
2.	Canning, preserving	149 9
3.	Baking	134 5
4.	Sugar, beet	93 4
5.	Slaughtering, meat packing	89 9
6.	Butter	87 3
7.	Confectionery	79 6
8.	Flour	75 5
9.	Ice cream	74 3
10.	Sugar refining, cane	74 1

Median = 88.6. Mean = 106.3

* Unit: per cent of 1923–1925 average. Source: *Federal Reserve Bulletin*, March, 1938, p. 228.

median. Where the variates near the median are scattered at considerable intervals (Table 54) the size of the median is somewhat accidental, depending on the value which happens to apply to the middle variate (or two middle variates).

The *geometric average* of a group of variates is the nth root of the product of those variates, where n is the number of variates. For all series with more than a very few variates the computation of the geometric average can best be done by the use of logarithms. Table 56 shows the logarithmic computation of the geometric average for Table 53. An intermediate step in the process is the

TABLE 55
DATA OF TABLE 53 ARRANGED IN AN ARRAY

Serial number	Index	Serial number	Index	Serial number	Index
32	11	19	73	9	101
17	16	20	74	23	108
34	42	35	74	25	108
6	43	18	76	28	108
2	49	16	78	10	114
1	50	37	84	30	123
4	54	26	87	15	134
14	63	3	88	29	140
5	66	12	88	13	142
7	69	24	88	11	143
33	70	31	89	36	247
22	71	21	93	27	255
8	72				

Median = 84

finding of the mean of the logarithms of the original variates. The logarithm of each given variate is found from a table of logarithms (see table beginning page 413); and then the mean of the logarithms is found by adding them and dividing by the number of variates. The antilogarithm of this mean logarithm is then found from the table of logarithms, and the result is the geometric average. The relation of the geometric average to the mean of the logarithms implies that some of the peculiar mathematical properties of the mean can be extended, with appropriate modifications, to the geometric average.

The *harmonic average* of a group of variates is the reciprocal of the mean of the reciprocals of the variates. Table 57 exhibits

the calculation of the harmonic average for the data of Table 53. Here again the fact that a mean enters as an intermediate step explains the partial extension of properties of the mean to the harmonic average.

Some further types of averages will be presented in the following chapter. Those listed above are the simple types most frequently used in summarizing categorical data. It has already

TABLE 56

CALCULATION OF THE GEOMETRIC AVERAGE, G, OF THE DATA OF TABLE 53

Serial number	Relative R	Logarithm of R	Serial number	Relative R	Logarithm of R
1	50	1.69897	20	74	1 86923
2	49	1 69020	21	93	1.96848
3	88	1 94448	22	71	1 85126
4	54	1 73239	23	108	2 03342
5	66	1 81954	24	88	1 94448
6	43	1 63347	25	108	2 03342
7	69	1 83885	26	87	1.93952
8	72	1 85733	27	255	2 40654
9	101	2 00432	28	108	2 03342
10	114	2 05690	29	140	2 14613
11	143	2 15534	30	123	2 08991
12	88	1 94448	31	89	1 94939
13	142	2 15229	32	11	1 04139
14	63	1 79934	33	70	1 84510
15	134	2.12710	34	42	1 62325
16	78	1 89209	35	74	1 86923
17	16	1 20412	36	247	2 39270
18	76	1 88081	37	84	1 92428
19	73	1 86332	Total		70 25649

$$\text{Mean log} = \frac{70\ 25649}{37} = 1.89882 = \log G.G = 79.22$$

been remarked that the mean most nearly meets the six requirements of Yule and Kendall, in ordinary problems. Each of the three other averages described above lacks one or more of the requisites: all three are less easily understood than the mean (requisite 3); the median is based only to a limited extent upon all the items (requisite 2), and is quite unfitted for algebraic manipulation (requisite 6), and neither the harmonic nor geometric average can be computed rapidly (requisite 4). The median is likely to be more typical than the mean in cases where there are several extremely large variates (Table 55); but, unless there are many

variates, the exact position of the median is erratic, as pointed out above for Table 54. Both the geometric and harmonic averages tend to reduce the effects of extremely large variates and to emphasize the small variates, the harmonic more than the geometric: thus for Table 53 the three computed averages have been found, in order of size, as harmonic (63.30), geometric (79.217) and mean (92). Whereas the mean is a magnitude such that the sum

TABLE 57

CALCULATION OF THE HARMONIC AVERAGE, H, OF THE DATA OF TABLE 53

Serial number	Relative R	Reciprocal of R	Serial number	Relative R	Reciprocal of R
1	50	0 02000	20	74	0 01351
2	49	0 02041	21	93	0 01075
3	88	0.01136	22	71	0 01408
4	54	0 01852	23	108	0 00926
5	66	0 01515	24	88	0 01136
6	43	0 02326	25	108	0.00926
7	69	0 01449	26	87	0 01149
8	72	0 01389	27	255	0 00392
9	101	0 00990	28	108	0 00926
10	114	0 00877	29	140	0.00714
11	143	0 00699	30	123	0 00813
12	88	0 01136	31	89	0 01124
13	142	0 00704	32	11	0.09091
14	63	0 01587	33	70	0 01429
15	134	0 00746	34	42	0.02381
16	78	0 01282	35	74	0.01351
17	16	0 06250	36	247	0 00405
18	76	0 01316	37	84	0 01190
19	73	0 01370	Total	.	0 58452

$$\text{Mean of reciprocals} = \frac{0.58452}{37} = 0.015798 = \frac{1}{H}. \quad H = 63.30$$

of the differences of the actual variates (values of the variable) from it is zero, the geometric average is a magnitude such that the product of the ratios of the variates to it is unity. This difference, really a difference in definition, indicates that in summarizing a series by use of the mean, emphasis is on *absolute* variations among the variates, whereas the geometric average emphasizes *relative* variations.

An artificial device frequently used in calculating averages is *weighting*. If certain variates are regarded as more important than others and should have greater influence upon the resulting

average, such variates can be given greater weight. This consists in *counting* such variates more than once in deriving the average. Thus for getting the weighted mean of the variates of column 1

TABLE 58

CALCULATION OF A WEIGHTED MEAN OF A CATEGORICAL SERIES, WITH INTEGRAL WEIGHTS*

Company	Earnings per share, 1937 E	Weight W	E entered W times	$W \cdot E$
	(1)	(2)	(3)	(4)
Addressograph-Multigraph	2.75	2	2.75 2.75	5.50
Burroughs Adding Machine	1 75	11	1.75 1.75 1.75 1 75 1.75 1.75 1.75 1 75 1.75 1.75 1 75	19.25
General Fireproofing	3 50	1	3.50	3 50
International Business Machines	11.50	2	11.50 11.50	23.00
National Cash Register	2.30	4	2.30 2 30 2 30 2.30	9 20
Pitney-Bowes Postage Meter	0.75	2	0 75 0 75	1 50
Remington Rand	2.75[a]	3	2 75 2 75 2 75	8 25
Telautograph	0.60	1	0 60	0.60
Underwood Elliott Fisher	6.50	2	6 50 6 50	13.00
Total	32 40	28	83.80	83 80

$$\text{Mean} = \frac{83\ 80}{28} = 2.99$$

* Unit, for earnings, one dollar per common share Source: Standard Statistics Company's *Earnings Bulletin*, January, 1938, p. 13, for earnings estimates. The weights are arbitrarily assigned, and are roughly proportional to the number of common shares outstanding, as given in Standard Statistics Company's *Basic Survey* for "Office and business equipment," March 3, 1937, pp. OF 12–OF 18. Two companies excluded because of incomplete earnings data.

[a] Fiscal year ending March, 1938.

of Table 58, with the weights as indicated in column 2, the computation might take the form shown in column 3. Clearly, a simplification results in replacing column 3 by column 4, wherein

each given variate is multiplied by its weight. This multiplication procedure admits the use of fractional as well as integral (whole number) weights (see Table 59). The divisor used in getting the weighted mean is not the *number* of variates, but the *sum* of the weights.

Weighting manifestly changes the value of the mean, in almost all cases. Thus the simple mean of the nine items of Table 58 is (sum of column 1, divided by 9) 3.60, whereas the weighted mean is 2.99. By exception, if the weights are related in a particular way

TABLE 59

CALCULATION OF THE WEIGHTED ARITHMETIC AVERAGE OF THE CATEGORICAL SERIES OF TABLE 54*

Symbol	Relative R	Weight W	$W \cdot R$
a. Baking	134.5	162 4	21,842 80
b. Beverages	204.9	27 9	5,716 71
c. Butter	87.3	19.0	1,658.70
d. Canning, preserving	149.9	83.6	12,531.64
e. Confectionery	79.6	62 4	4,967 04
f. Flour	75.5	33.5	2,529.25
g. Ice cream	74 3	23.3	1,731.19
h. Slaughtering, meat packing	89.9	126 7	11,390 33
i. Sugar, beet	93.4	8.2	765 88
j. Sugar refining, cane	74.1	14 7	1,089 27
Total		561 7	64,222 81

$$\text{Weighted mean} = \frac{64{,}222.81}{561.7} = 114.34$$

* Unit: for R, per cent of 1923–1925 average for W, thousand wage earners (average in 1923–1925 period). Source: for R, *Federal Reserve Bulletin*, March, 1938, p 228, for W, *ibid.*, December, 1936, p. 956.

to the variates, the two results might be identical. In general, if the smaller variates carry in the main the heavier weights, the weighted mean will be smaller than the simple mean, and *vice versa*.

The application of the weighting principle to the harmonic average is obvious. The weighted harmonic average is simply the reciprocal of the weighted arithmetic average of the reciprocals of the individual variates. For the weighted geometric average, the variate is entered in the "product of all the variates" a number of times equal to its weight: thus the weight of the variate becomes its exponent, in the formula representing the definition of geometric average. The index of the root taken is no longer n, the

number of variates, but is the sum of the weights (exponents). Table 60 shows the derivation of the weighted geometric average for the data of Table 59. A weighted median also can be obtained, with due care in cases of fractional weights; and, indeed, any type of average may be weighted. Further discussion of problems of weighting appears in Chap. XIX.

TABLE 60

CALCULATION OF THE WEIGHTED GEOMETRIC AVERAGE OF THE SERIES OF TABLE 59*

Symbol	Relative R	Logarithm of R	Weight W	$W \cdot \log R$
a	134 5	2 12872	162.4	345 704128
b	204.9	2 31154	27 9	64 491966
c	87.3	1 94101	19 0	36 879190
d	149 9	2 17580	83 6	181 896880
e	79.6	1 90091	62 4	118.616784
f	75 5	1 87795	33.5	62 911325
g	74 3	1 87099	23.3	43 594067
h	89 9	1 95376	126 7	247 541392
i	93 4	1 97035	8 2	16 156870
j	74 1	1 86982	14 7	27 486354
	Total	.....	561 7	1,145.278956

Mean log $= \frac{1{,}145.278956}{561.7} = 2.03895$. Antilog of same is geometric mean = 109.38

* For units and sources of data, see Table 59.

RATES AND RATIOS

We have already noted (Chap. II) a large class of summary statistical numbers called *rates* or *ratios*. These are quotients in which the numerator and denominator are generally less closely related than in any of the averages; although some of them, particularly the per capita ratios, are averages by implication. In a large number of other instances, moreover, the concept of the mean is implied in the definition of the particular rate or ratio: the numerator and denominator of the ratio often partake of the nature of the aggregate and the number of cases, respectively. This fact is especially significant because on this ground many such quotients can be examined to determine their fitness as typical summarizations of the data.

For a ratio to have real meaning as a statistical summarization, the numerator and denominator must have a significant relation to

each other. In some cases this relation exists because of the direct connection between numerator and denominator arising from their derivation from the data. In other cases the relation is not inherent in the data but arises from customary interpretation of the data. The essential point is that the value of a quotient can vary as a result of change either in the dividend (numerator) or the divisor, and the use of an improper divisor often introduces causes of variation which are misleading to the reader who interprets the ratio. Moreover, as indicated above, many of these ratios are in fact averages; and, as such, they should receive the same scrutiny as is given to averages. This really involves, as shown in the following chapter, knowledge of the frequency distribution which lies behind the ratio. A ratio can be nontypical in the same way as an average; and this flaw is generally less easily discovered in the case of the ratio than for an average, partly because the corresponding frequency data may not be available and partly because the definition of the ratio may be so intricate that we find difficulty in representing the situation in frequency form.

CHAPTER XI

AVERAGES FOR FREQUENCY SERIES

AVERAGES OF DISCRETE FREQUENCY SERIES

The averages defined in the preceding chapter are obtainable when the data appear in the form of a frequency series as well as for the categorical data there used in illustration of the calculations. In fact, the theoretical aspects of the problem of averages and the proper limitations upon the use of averages can best be studied in connection with frequency distributions. The actual work of calculation is generally more arduous for frequency data, and assumptions or approximations must often be made which are not essential in the analysis of categorical data.

A basic peculiarity of the frequency series is that the item is not a variate, as in a categorical or time series, but is the number of observed objects (or instances) for which the variable has a specified size or for which the variable falls within a specified size interval. The average pertains to the *variable;* and therefore, whereas the average is derived from the *items* in a categorical or time series, the average of a frequency series is a value of the variable magnitude usually appearing in the *stub.*

Table 61 presents a categorical series in which the variable is capitalization, and each item is the capitalization of a particular bank. According to the definitions of the preceding chapter, the several averages of this series can be calculated. Thus the mean capitalization per bank is 191.64 thousand dollars, the quotient of the total of all the items by the number of items.

These same data can be classified according to size, yielding the frequency series shown in the left column of Table 62. This series is discrete, for each element of the rule of classification is a specified size rather than a size interval. As these data are precisely equivalent to those of Table 61, having merely a different form of presentation, the averages computed for this frequency series must be identical with those found for the categorical series in Table 61.

The mean for Table 62 is found by weighting each specified size by the number of banks with capitalization of that size. In

other words, each value of the variable X is multiplied by the corresponding frequency f, these products are added, and the sum is divided by the total frequency. In actual practice, these operations are performed by means of a working table designed like

TABLE 61

CAPITAL OF NATIONAL BANKS IN MICHIGAN ON DECEMBER 31, 1936*

Serial number	Capital[a]	Serial number	Capital[a]	Serial number	Capital[a]
1	110	32	100	5	150
2	910	33	100	6	50
3	540	34	37 5	7	500
4	988	35	100	8	165
5	400	36	210	9	25
6	142.5	37	600	10	300
7	100	38	550	11	50
8	50	39	100	12	100
9	50	40	100	13	62 5
10	100	41	50	14	100
11	100	42	490	15	100
12	165	43	1000	16	50
15	50	44	50	17	50
16	50	1	50	18	400
17	35	2	50	19	150
18	460	3	1250	20	200
21	50	4	50	21	100
22	50	5	50	22	100
23	50	6	50	23	100
24	100	7	50	24	75
25	35	8	50	25	50
26	290	9	122	26	37 5
27	650	10	100	27	50
28	250	1	45	28	275
29	350	2	50	29	80
				Total	15,139 8
30	164.8	3	50	Average	191 64
31	25	4	100		

* Source. "Individual Statements of Condition of National Banks at Close of Business, December 31, 1936" (Table N, issued as supplement to Annual Report of Comptroller of the Currency), Washington, 1937, pp. 65–67. The three sequences of serial numbers are as in the source. Two banks each in the reserve cities, Detroit and Grand Rapids, have been excluded.
[a] Unit: thousand dollars.

Table 62 but perhaps less formal in arrangement, although all essential features and the orderly arrangement of Table 62 would exist in a good working table. It should be observed that this process is exactly equivalent to that followed for the categorical series: the capital of each bank is entered once in the aggregate, and the aggregate is divided by the total number of banks. Here,

TABLE 62

DISCRETE FREQUENCY SERIES DERIVED FROM TABLE 61, WITH DIRECT CALCULATION OF MEAN

Capital[a] X	Frequency f	fX
25	2	50
35	2	70
37 5	2	75
45	1	45
50	24	1200
62.5	1	62.5
75	1	75
80	1	80
100	17	1700
110	1	110
122	1	122
142.5	1	142.5
150	2	300
164.8	1	164 8
165	2	330
200	1	200
210	1	210
250	1	250
275	1	275
290	1	290
300	1	300
350	1	350
400	2	800
460	1	460
490	1	490
500	1	500
540	1	540
550	1	550
600	1	600
650	1	650
910	1	910
988	1	988
1000	1	1000
1250	1	1250
Total	79	15,139.8

$$\text{Mean} = \frac{15{,}139.8}{79} = 191.64 \text{ units, or } \$191{,}640$$

[a] Unit: thousand dollars.

TABLE 63

CALCULATION OF THE HARMONIC AVERAGE FOR THE DISCRETE SERIES OF TABLES 62 AND 46

Capital[a] X	Frequency f	$\frac{1}{X}$	$\frac{f}{X}$
25	2	0 040000	0 080000
35	2	0 028571	0 057142
37 5	2	0 026667	0 053334
45	1	0.022222	0.022222
50	24	0.020000	0 480000
62.5	1	0 016000	0.016000
75	1	0 013333	0.013333
80	1	0.012500	0.012500
100	17	0.010000	0.170000
110	1	0.009091	0.009091
122	1	0.008197	0.008197
142 5	1	0 007018	0 007018
150	2	0.006667	0 013334
164 8	1	0.006068	0 006068
165	2	0.006061	0.012122
200	1	0 005000	0.005000
210	1	0.004762	0 004762
250	1	0.004000	0.004000
275	1	0 003636	0.003636
290	1	0.003448	0.003448
300	1	0 003333	0 003333
350	1	0.002857	0 002857
400	2	0 002500	0 005000
460	1	0 002174	0 002174
490	1	0 002041	0 002041
500	1	0 002000	0.002000
540	1	0 001852	0 001852
550	1	0 001818	0 001818
600	1	0 001667	0 001667
650	1	0.001538	0.001538
910	1	0 001099	0 001099
988	1	0 001012	0.001012
1000	1	0 001000	0.001000
1250	1	0 000800	0.000800
Total	79		1 009398

$$\text{Harmonic average} = \frac{79}{1.009398} = 78.264 \text{ units, or } \$78{,}264$$

[a] Unit: thousand dollars.

TABLE 64

CALCULATION OF THE GEOMETRIC AVERAGE FOR THE DISCRETE SERIES OF TABLE 62

Capital[a] X	Frequency f	log X	f log X
25	2	1.3979	2 7958
35	2	1 5441	3.0882
37.5	2	1 5740	3.1480
45	1	1.6532	1.6532
50	24	1 6990	40.7760
62.5	1	1.7959	1.7959
75	1	1.8751	1.8751
80	1	1 9031	1.9031
100	17	2 0000	34.0000
110	1	2 0414	2.0414
122	1	2.0864	2.0864
142 5	1	2 1538	2 1538
150	2	2 1761	4.3522
164 8	1	2 2170	2.2170
165	2	2 2175	4.4350
200	1	2.3010	2 3010
210	1	2.3222	2.3222
250	1	2.3979	2.3979
275	1	2 4393	2 4393
290	1	2 4624	2.4624
300	1	2.4771	2.4771
350	1	2.5441	2.5441
400	2	2 6021	5.2042
460	1	2 6628	2 6628
490	1	2 6902	2.6902
500	1	2.6990	2.6990
540	1	2 7324	2.7324
550	1	2 7404	2.7404
600	1	2.7782	2.7782
650	1	2.8129	2.8129
910	1	2 9590	2.9590
988	1	2 9948	2 9948
1000	1	3 0000	3 0000
1250	1	3 0969	3 0969
Total	79		161 6359

$$\log G = \frac{161.6359}{79} = 2.0460. \quad G = 111.18 \text{ units, or } \$111{,}180$$

[a] Unit: thousand dollars.

TABLE 65

LOCATION OF THE MEDIAN OF THE DISCRETE SERIES OF TABLE 62

Capital[a] X	Frequency f	Cumulative frequency	
		Equal to or less than X	Equal to or greater than X
25	2	2	79
35	2	4	77
37 5	2	6	75
45	1	7	73
50	24	31	72
62 5	1	32	48
75	1	33	47
80	1	34	46
100	17	51	45
110	1	52	28
122	1	53	27
142 5	1	54	26
150	2	56	25
164.8	1	57	23
165	2	59	22
200	1	60	20
210	1	61	19
250	1	62	18
275	1	63	17
290	1	64	16
300	1	65	15
350	1	66	14
400	2	68	13
460	1	69	11
490	1	70	10
500	1	71	9
540	1	72	8
550	1	73	7
600	1	74	6
650	1	75	5
910	1	76	4
988	1	77	3
1000	1	78	2
1250	1	79	1
Total	79		

Median is item number $\frac{79+1}{2}$, or the 40th item, which = \$100,000

[a] Unit: thousand dollars.

however, all banks with equal capital are taken in one group instead of being added singly according to the original order of Table 61.

The finding of the harmonic and geometric averages for a discrete frequency series rests upon similar adaptations of the methods used for categorical series. Tables 63 and 64 are the working forms for these calculations, based upon the frequency data of the first column of Table 62. The median of the frequency series can be found more easily than for the corresponding categorical series because the variates in a frequency series are already arrayed. One usually derives the cumulative frequencies, and the location of the middle variate is then obvious. Thus, as in Table 65 the total frequency is 79, the middle variate is the $\frac{79+1}{2}$th; and this obviously lies in the group having 51 as its downward cumulative frequency. Hence the median is $100,000.

THE MEAN OF A GROUPED FREQUENCY SERIES

Although the data of Table 61 can be presented in the form of a grouped frequency series, such presentation is not appropriate, because the variable in this case is known to be discontinuous and to have frequencies concentrated at particular values. It is possible to capitalize a bank for any amount, within statutory limits, but custom dictates that the capital value be in round numbers. The forcing of such data, which pertain to an intrinsically discrete variable, into the form of a grouped (continuous) frequency series is never statistically accurate and is seldom justifiable. The present instance is one in which such forcing would lead to falsification of the facts. Allocation of variates to class *intervals* would yield frequencies for which assumptions made below in calculating averages and other characteristics of grouped frequency series would not be valid. The arithmetical analysis of such grouped series, improperly formed from categorical data representing an essentially discontinuous variable, would result in significant errors.

In illustration of this objectionable practice, Table 66 is presented. Here the "grouping" of these freight-car capacities is misleading; the actual size (as indicated by the stated average for each group) of the large majority of the cars falling within any one interval, with few exceptions, lies very near the lower limit of that interval. The specific bearing of this situation upon the computed results of the averages and other summary numbers will appear

later. On the other hand, grouping data pertaining to a truly discrete variable in a continuous frequency series is a very common practice and is often quite justifiable. Careful examination will show that in these acceptable instances two conditions ordinarily hold: The total frequency is large; and the intervals between actual values of the discrete variable are small as compared with the class interval, and not highly irregular in width.

No real difficulty arises for the series which is only *formally* discrete: the series in which the variable phenomenon is con-

TABLE 66

NUMBER OF FREIGHT CARS HAVING TONNAGE CAPACITY WITHIN EACH SPECIFIED INTERVAL, AND THE AVERAGE CAPACITY FOR EACH INTERVAL, FOR RAILWAYS OF ALL CLASSES ON DECEMBER 31, 1916*

Capacity[a]	Number of cars	Average capacity
Below 10	1,547	6.0
10 –15	4,102	11.2
15– 20	3,105	15.2
20– 25	29,046	20.1
25– 30	46,795	25.0
30– 35	614,238	30.1
35– 40	39,318	35.0
40– 45	764,496	40.1
45– 50	9,136	45 0
50 –55	644,004	50 1
55– 60	140,450	55 4
60– 65	1,184	60 0
65– 75	29,432	69 9
75–100	1,988	87 9
100 and over	380	100 2

* Source: "Statistics of Railways in the United States, 1916," Washington, Interstate Commerce Commission, 1919, pp. 18–21.
[a] Upper limit of class interval excluded. Unit: one ton.

tinuous, but for which the measuring devices are such that the measurements appear discontinuous. A similar formal discontinuity arises from "rounding off" measurements. Length, weight, price, temperature, the rate of interest, and in fact nearly all variable magnitudes which are subject to measurement, yield series of this sort. In such cases, however, unless the measuring instruments are exceptionally faulty, no objection exists to treating the data as continuous.

The variable of Table 67 is essentially continuous, as the earnings of an individual bookkeeper can vary by amounts as small as can be measured—subject to the limitations of custom which

states wages usually in round figures. The corresponding frequency series given in the table may therefore be regarded as continuous. The class intervals are of uniform width, and there are no "all over" or "all under" groups. These features are desirable, if not essential, in order that the analysis of the series may proceed promptly and lead to results relatively free from arbitrary assumptions. An essential preliminary step in any

TABLE 67

NUMBER OF MACHINE BOOKKEEPERS, IN A SAMPLE OF NEW YORK CLERICAL WORKERS IN MAY, 1937, HAVING SPECIFIED WEEKLY EARNINGS, AND CALCULATION OF THE MEAN*

Weekly earnings X	Number of bookkeepers f	fX
17 5	88	1540 0
22 5	229	5152 5
27 5	372	10230 0
32 5	139	4517 5
37 5	54	2025 0
42 5	14	595 0
47 5	10	475 0
52 5	15	787 5
57 5	1	57 5
62 5	2	125 0
67 5	1	67 5
72 5	1	72 5
Total	926	25645 0

$$\text{Mean} = \frac{25645.0}{926} = 27.69 \text{ units, or } \$27.69$$

* Data from Table 40, and source there given. Earnings entries are in dollars, for center of each class interval

thorough analysis of a continuous frequency series is the graphing of the data, and Chart 40 shows the frequency polygon.

The computation of the mean of a grouped frequency series involves an assumption concerning the distribution of frequency within any particular class interval. The assumption made is that the mean of all values of the variable falling within a given class interval is exactly at the center of that interval (that is, at the class mark). Although this assumption is such that it would correctly describe the distribution within the interval for a large number of widely differing types of such distribution, there are many actual series in which even this relatively light requirement

is fulfilled only approximately, and for which the assumption therefore introduces an error. This assumption once made, the series can, for the purpose of computing the mean, be treated as a discrete series. The computation of the mean then amounts to getting the weighted average of the several class marks, the weights being the stated frequencies. For Table 67 each mid-value of column 1—the mid-values alone are stated in the stubs, and not the class limits—is multiplied by the adjacent item of column 2. The products are summed and divided by 926 to yield the mean.

THE SKELETON METHOD OF COMPUTATION

For series in which the variable ranges over a wide succession of values, and for which the frequencies are large, the arithmetic of the above process may well become burdensome. Two special devices are used to reduce the labor of computation: The *unit* of measurement is changed from the natural unit (one dollar in this case) to the width of one class interval; and the *origin* of measurement is changed from the natural origin (zero dollars, in this case) to an arbitrary origin, usually selected at a class mark at or near which the mean is expected to fall.

The result is to express the *variable* in a new form, indicated by the numbers x of column 3 of Table 68. Then the original frequencies f (column 2) are considered as forming, with these "variables" x, a new discrete frequency series. The products xf are entered in columns 4 and 5, and the net total of these columns is divided by the total frequency to yield the weighted mean of the numbers x, that is, the mean of this artificial frequency series.

It remains to convert this result, $d = 1.0389$, to a form which can be accepted as the mean of the original series. This involves changing back to the original units and then to the original origin, by adding 5.1945 (which is 5×1.0389) to 22.5. The result, 27.69, is the mean, and is of course identical with that found by use of column 6 and the related columns 7 and 8. This result may well differ, however, from that which would be found for the original categorical series if it were available; and this difference indicates the extent of the error arising from assuming the frequency in each class interval centered at the class mark.

This error is in no way due to the skeleton method but results from the grouping of the frequencies in classes and from assuming the class frequencies distributed uniformly so that the mean of the variates in a class is at the center of that interval (as in Table 67). In actual practice, to be sure, the computation from the categori-

cal data would not have been made, if indeed such data were available (and available data often come, as in this case, only in the form of frequency series). It is therefore important to note that the error introduced by the above assumption is not likely to be large, unless the series is markedly irregular or unsymmetrical (which would appear from the chart) or the grouping has been made on some inappropriate basis.

TABLE 68

CALCULATION OF THE MEAN FOR TABLE 67 BY THE SKELETON METHOD, WITH CHECK COMPUTATION

Weekly earnings X	Number of bookkeepers f	x	xf		x'	$x'f$	
			−	+		−	+
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
15–20	88	− 1	−88		−3	− 264	
20–25	229	0			−2	− 458	
25–30	372	1		372	−1	− 372	
30–35	139	2		278	0		
35–40	54	3		162	1		54
40–45	14	4		56	2		28
45–50	10	5		50	3		30
50–55	15	6		90	4		60
55–60	1	7		7	5		5
60–65	2	8		16	6		12
65–70	1	9		9	7		7
70–75	1	10		10	8		8
Total	926		−88	+1050		−1094	+204
				+962			−890

$d = {}^{96}\!/_{926} = 1.0389$ class intervals. $d' = -{}^{890}\!/_{926} = -0.9611$ class intervals
Mean = 22.5 + 5.1945 = 27.69 units, or Mean = 32.5 − 4.8055 = 27.69 units

The method of the arbitrary origin and special unit, hereafter called the *skeleton method*, may not appear advantageous for the problem used in the illustration. For series of large frequency, however, its comparative speed and simplicity are obvious; and its use in the evaluation of characteristics other than the mean leads to greater economies in time and effort, even for series in which the total frequencies are only moderately large. The choice of the center of some other class interval (column 6 of Table 68) alters all the intermediate results, but the final value of the mean is identical with that already found. This indicates that a very effective *check* upon the computation is obtainable by selecting two different arbitrary origins. It should be noted also that the skeleton

method is not fully applicable in any series in which the width of the class intervals varies (Table 69). The arbitrary origin in such a case can be chosen; but although a new unit commensurable with all the class intervals can be adopted, the searching for such a unit is seldom repaid by the saving in subsequent computing effort, for not all class intervals have unit width. In Table 69 the final columns give, for the broader intervals, the average frequency per $10,000 subinterval within each broad interval. In applying the skeleton method, for Table 69, fifteen subintervals $10,000

TABLE 69

SALARIES, WAGES, ETC., REPORTED ON INDIVIDUAL RETURNS OF NET INCOME OF $5,000 OR OVER FOR UNITED STATES INCOME TAX, 1934*

Lower limit of net income ($1,000)	Number of returns	Amount ($1,000)	Adjustment factor	Adjusted frequency[b]	
				Number	Amount ($1,000)
0	209,254	1,173,650	1	209,254	1,173,650
10	48,937	644,432	1	48,937	644,432
20	8,921	211,384	1	8,921	211,384
30	2,754	93,152	1	2,754	93,152
40	1,169	51,669	1	1,169	51,669
50	1,155	67,866	2.5	462	27,146
75	287	24,336	2.5	114 8	9,734 4
100	215	28,364	15.	14 3	1,890.9
250	10	3,671	25.	0 4	146 8
500	3	1,732	50.	0.06	34 6
1000	0	0	[a]	.....	.....
Total	272,705	2,300,256	....	..	.....

* Source: "Statistics of Income, 1934, Part 1," Washington, U. S. Treasury, 1936, p. 13. (Certain combinations have been made of data as given.)

[a] No figure possible, because upper limit not known.

[b] Given figures divided by adjustment factor.

wide would be inserted in the class interval 100–250 and fifteen successive integral values of x would be assigned to the centers of these subintervals, and a frequency of 14.3 (using the series for "number of returns") would pertain to each. A difficulty would arise in the fractional cases: the half subinterval 70–75 presumably includes 231 cases, and 75–80 includes 57.4 cases, so that the $10,000 subinterval 70–80, with an x chosen for its center 75, would be assigned a frequency of 288.4. Obviously, these operations involve dubious assumptions concerning allocation of frequencies to subintervals. The alternative procedure is to calculate x in decimal fractions for the center of each given interval

TABLE 70

DISTRIBUTION OF EMPLOYEES IN QUARRIES IN DIMENSION-GRANITE INDUSTRY CLASSIFIED ACCORDING TO WEEKLY EARNINGS, AUGUST, 1937*

Weekly earnings[a]	Number of employees			
	Total	Skilled	Semiskilled[b]	Unskilled
Under 5	28	14	5	9
5–10	137	64	29	44
10–15	200	91	42	67
15–20	258	125	63	70
20–25	477	350	56	71
25–30	376	326	36	14
30–35	127	112	11	4
35–40	63	59	4	0
40–45	39	39	0	0
45–50	9	9	0	0
50 and over	27	27	0	0

* Source *Monthly Labor Review*, December, 1937, pp. 1487, 1489.
[a] Lower limit inclusive. Unit: one dollar.
[b] Includes apprentices and learners

TABLE 71

DIVIDENDS ON STOCK OF DOMESTIC CORPORATIONS REPORTED ON INDIVIDUAL RETURNS OF NET INCOME OF $5,000 AND OVER FOR UNITED STATES INCOME TAX, 1934*

Lower limit of net income ($1,000)	Number of returns	Amount ($1,000)	Adjustment factor	Adjusted frequency[b]	
				Number	Amount ($1,000)
0	231,299	456,198	1	231,299	456,198
10	19,045	263,457	1	19,045	263,457
20	6,392	154,733	1	6,392	154,733
30	3,019	104,003	1	3,019	104,003
40	1,656	73,731	1	1,656	73,731
50	1,927	115,829	2 5	770 8	46,331.6
75	802	68,668	2 5	320 8	27,467 2
100	1,008	149,922	15	67 2	9,994 8
250	219	72,037	25.	8 8	2,881 5
500	85	58,984	50.	1.7	1,179.7
1000	35	67,482	[a]	.....	.....
Total	265,487	1,585,043[c]	....	.....	.....

* Source: "Statistics of Income, 1934, Part I," Washington, U. S. Treasury, 1936, p. 14. (Certain combinations have been made of data as given.)
[a] No figure possible, because upper limit not known.
[b] Given figures divided by adjustment factor.
[c] This total figure, as given in the source, is $1,000 smaller than the sum of the individual items.

and weight it by the *given* frequency. This is simpler for the mean calculation but results in burdensome labor for other characteristics of the series. Regardless of the method of calculation, the series of Table 69 are so very asymmetrical that the mean is a poor average (see below, page 225). Series with "all under" or "all over" groups present difficulties in any method: some assumption, usually very uncertain, must be made for "centering" the frequencies in such intervals (Table 70). Table 71 illustrates series having both defects, intervals of unequal width and an all-over interval of uncertain width.

PROPERTIES OF THE MEAN

As noted in Chap. X (page 162), the mean possesses to a marked degree the six general requisites of a good average. It is in the

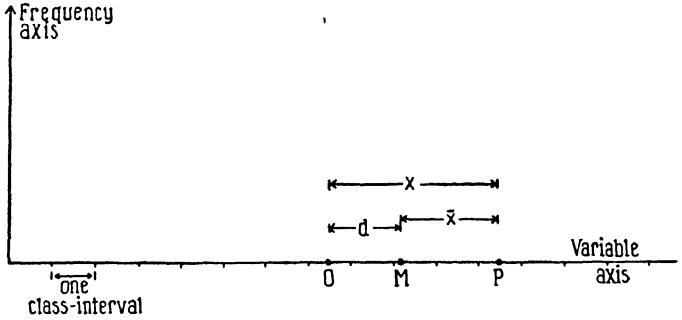


CHART 73.—Relation between deviations from the mean, and from the arbitrary origin.

sixth requisite, suitableness for algebraic manipulation, that the mean has an especial advantage over many other types of average. The following three outstanding algebraic properties of the mean should be carefully noted.

The mean is an invariant characteristic of the series: its location is independent of the choice of origin or unit of measurement. Although the *numerical* result differs with different origins or units, the position of the mean, at a definite point in a definite class interval, does not change. The skeleton method of computation depends upon this property.

The mean is such a value of the variable that the sum of the deviations (with due regard for plus and minus signs) of the actual variates from it is zero. This property becomes evident upon consideration of a hypothetical case in which the arbitrary origin happens to be selected precisely at the mean. Obviously, under

such conditions, the resulting d, which is the sum of the deviations of the several variates from the mean divided by the total frequency, must be zero.

The mean is such a value of the variable that the sum of the *squares* of the deviations of the actual variates from it is less than the sum of the squares of such deviations from any other value of the variable.[1] This third property of the mean is a form of the *principle of least squares:* the mean is a value of the variable such that the sum of the squares of deviations from it is least. This last statement may, in fact, be taken as a definition of the mean. Important consequences of this feature of the mean will appear later (Chap. XIII).

THE MEDIAN OF A GROUPED FREQUENCY SERIES

The determination of the class interval in which the median lies follows the same rule as the location of the median of a discrete series. The median is the $\frac{n+1}{2}$th variate where n is the total frequency, and such variate lies in the class interval for which the cumulative frequency exceeds (or is as large as) $\frac{n+1}{2}$ and for

---

[1] To establish this principle, let the deviation from some other point (say the arbitrary origin) be expressed in terms of the deviation from the mean. Thus in Chart 73, x is the deviation of the center, P, of a particular class interval from the arbitrary origin o; $\bar{x}$ is the deviation of P from the mean, M; and d is the distance of M from o (all measurements made in terms of some single unit, of any convenient length). Then, obviously, $x = \bar{x} + d$; and the square of the deviation from o is

$$x^2 = \bar{x}^2 + 2\bar{x}d + d^2$$

and the sum of the squares of the deviations from o is

$$\Sigma x^2 f_x = \Sigma(\bar{x}^2 + 2\bar{x}d + d^2)f_x$$

where f_x is the frequency in the class interval of which P is center, and the symbol Σ implies "the sum of all terms like," one such term for every value of x covered by the given series. The right-hand expression above can be broken into three parts;

$$\Sigma \bar{x}^2 f_x + \Sigma 2\bar{x}d f_x + \Sigma d^2 f_x$$

and, as d is the same for all terms in the summation, this becomes

$$\Sigma \bar{x}^2 f_x + 2d\Sigma \bar{x} f_x + d^2 \Sigma f_x$$

The first summation here is clearly the sum of the squares of the deviations from the mean; the second summation is zero, because it is the sum of the deviations from the mean and was shown above in the text to be zero; and the last summation is the total frequency N. It follows that the sum of the squares of the deviations from o is equal to the sum of the squares of the deviations from M plus d^2N. As d^2N is necessarily positive, the property is demonstrated: the minimum value of the sum is attained when d is zero.

which the next smaller cumulative frequency (for the adjacent interval) is less than $\frac{n+1}{2}$. The location of the median at a specific point in such class interval requires an assumption concerning the nature of the distribution of frequency within the interval, and this assumption is more restrictive than that made in calculating the mean (see above, page 176). The assumption ordinarily made for getting the median is a very special case under the assumption made in getting the mean. It is assumed that, if there are f_x variates within the interval, and if the interval is divided into f_x equal subintervals, one variate falls precisely at the center of each subinterval. Thus, if the width of the class

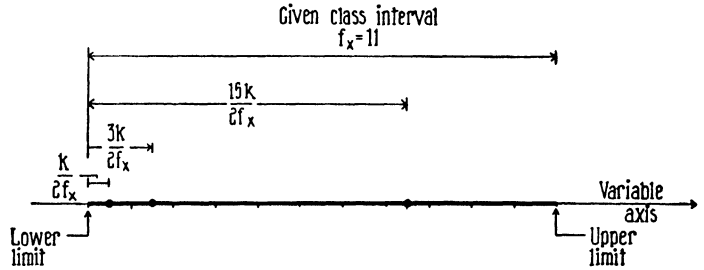


CHART 74.—Illustration of assumed equal spacing of the variates within a class interval.

interval is k, the lowest variate in the interval is at a distance $k/2f_x$ above the lower limit of the interval, the next variate $3k/2f_x$ above the lower limit, and so on (Chart 74). With this assumption, the median can be calculated by the notion of proportionality, as indicated in the calculations appended to Table 72. By cumulating the frequencies from the other end of the scale, column 3, an independent value of the median can be found, and this value should check the first result.

Class intervals of differing widths offer no obstacles to finding the median. Moreover, the interval "all under" or "all over" has no effect upon the process, barring the highly unusual case in which more than half the total frequency falls in such interval. It is also significant that the proportionate distribution assumption, although severely specific for the particular interval in which the median falls, does not extend to the other intervals. The form of the distribution of frequency within all other intervals is immaterial to the computation; and an important fact is that in practice

the assumption is likely to be very nearly correct for the limited range of the variable in the vicinity of the median, whenever that interval is narrow and near the mode.

TABLE 72
CALCULATION OF MEDIAN FOR THE SERIES OF TABLE 67

Weekly earnings X	Ordinary frequency	Cumulative frequency	
		Equal to or less than upper limit	Equal to or greater than lower limit
	(1)	(2)	(3)
15–20	88	88	926
20–25	229	317	838
25–30	372	689	609
30–35	139	828	237
35–40	54	882	98
40–45	14	896	44
45–50	10	906	30
50–55	15	921	20
55–60	1	922	5
60–65	2	924	4
65–70	1	925	2
70–75	1	926	1
Total	926		

Median $= 25 + 146 \times \frac{5}{372} = 25 + 1.96 = 26.96$ units,

or

Median $= 30 - 226 \times \frac{5}{372} = 30 - 3.04 = 26.96$ units

PROPERTIES OF THE MEDIAN

The median does not possess the six fundamental requisites (page 157) of a good average to such marked degree as does the mean. An especially serious drawback to the median is the great difficulty incident to its algebraic treatment. A limitation upon its significance as a typical value of the variable is that the large majority of the actual variates have only slight influence upon it. It is the *position* of these variates above or below the median, rather than their specific size, which counts in determining the median. This fact is an advantage in that it protects the median from the influence of irregularly large or irregularly small variates, and that it renders the calculation of the median possible for series in which the specific size of outlying variates is unknown. It is, on the other hand, a serious disadvantage in that the value of

the median is independent of the precise sizes of large numbers of variates moderately close to but on either side of the median.

An algebraic property of the median which assists in describing its relation to the series as a whole is that the median is a value of the variable such that the sum of the absolute (irrespective of plus and minus signs) deviations from it is a minimum.[1]

THE MODE OF A GROUPED FREQUENCY SERIES

The *mode* of a frequency series is that value of the variable for which the frequency is a maximum. In a distribution of simple form and free from irregularities, the approximate value of the mode can be estimated from the curve, especially in the form of a block diagram (Chart 76). If irregularities are present, as is ordinarily true in practice, the graphic estimate of the mode is made from the smoothed curve (Chart 77). The numerical

[1] This becomes apparent from Chart 75. The total frequency is assumed even, but analogous steps would constitute the argument for an odd total frequency. The median C lies midway between the two central variates, indicated by spots on the horizontal axis. Let o be an arbitrary point, also between those two variates. The sum of the absolute deviations of the variates from o equals the sum of such deviations from C because, for the $n/2$ smaller variates, each of the deviations from o is h greater than that from C, and for the $n/2$ larger variates, each of the deviations from o is h less than from C, and hence the aggregate is the same for o as for C. Consider next o', outside the interval between the two central variates. Here that portion of the aggregate contributed by deviations of the $(n/2) - 1$ variates below the central pair and of the $(n/2) - 1$ variates above the central pair from o' is the same as the portion contributed by corresponding deviations from C. The absolute deviations of the two central variates from o', however, total $2a$ more than those from C. Hence the property is established.

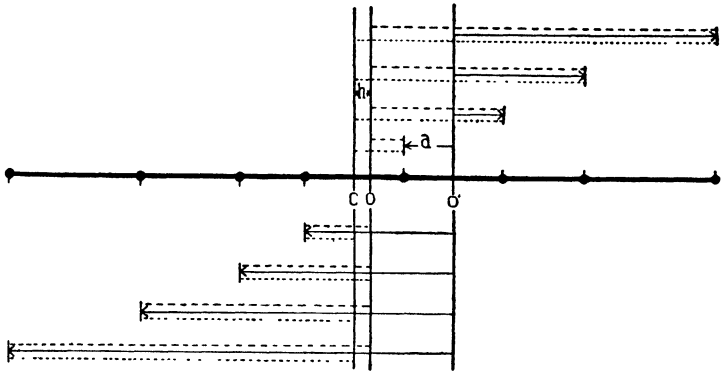


CHART 75.—Deviations of variates from the median and from arbitrary points on either side thereof.

evaluation of the mode for such series implies precise "fitting" of a mathematical curve, but approximations are possible by use of a formula derived from relations revealed in the "fitting" process, or by successive regrouping of frequencies.

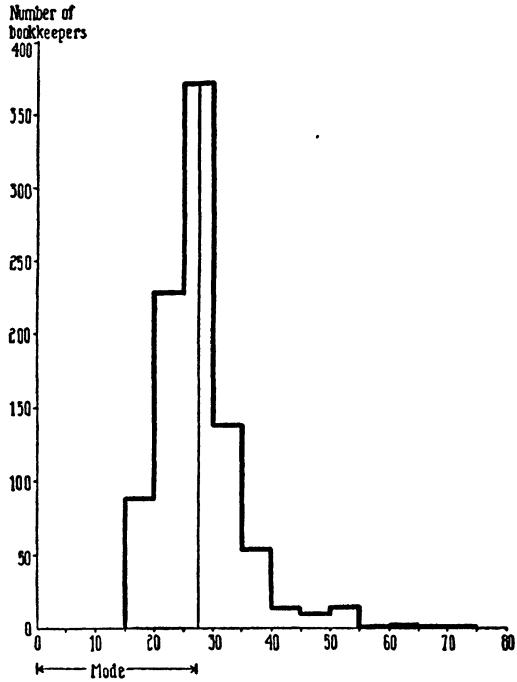


CHART 76.—Location of apparent mode from a block diagram—case of number of bookkeepers in New York City in May, 1937, classified by weekly earnings. (Data in Table 67, page 176.)

The formula, which is applicable only for series which are unimodal and only moderately asymmetrical (Chart 77, but not Chart 78 or Chart 79), is:

$$Mode = 3\ Median - 2\ Mean$$

This formula estimates the *true* mode, the mode as it would appear from a smooth curve properly fitted to the series, whereas the mode found from Chart 76 is the *apparent* mode—the variate for which the actual frequency is maximum.

In a particular case, one examines the curve to be sure that it is of the type for which the formula is appropriate, then calculates

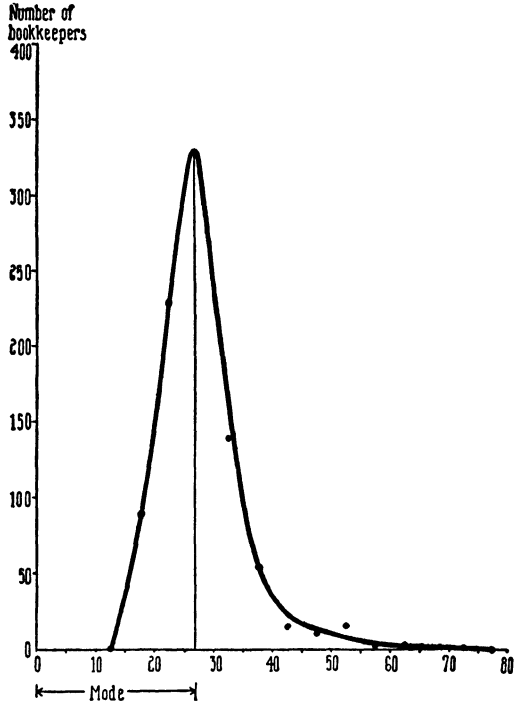


CHART 77.—Estimate of true mode of series of Chart 76, by fitting smooth curve visually to plotted points.

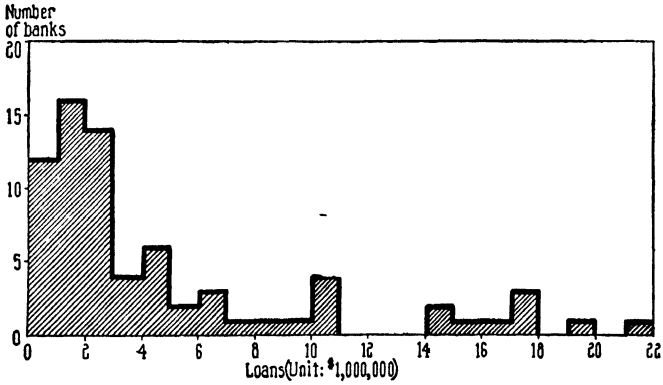


CHART 78.—Number of Michigan national banks having loans within stated ranges on December 31, 1936.

(Data in Table S, Appendix A.)

mean and median, and derives the mode by direct substitution in the formula. The result is, of course, only approximate. The formula may be written also as follows:

$$Mode = Mean - 3(Mean - Median)$$

and this form of statement exhibits the order of magnitude of the three averages as ordinarily found in a moderately asymmetrical curve. The median lies between the mean and mode, and twice as far from the mode as from the mean.

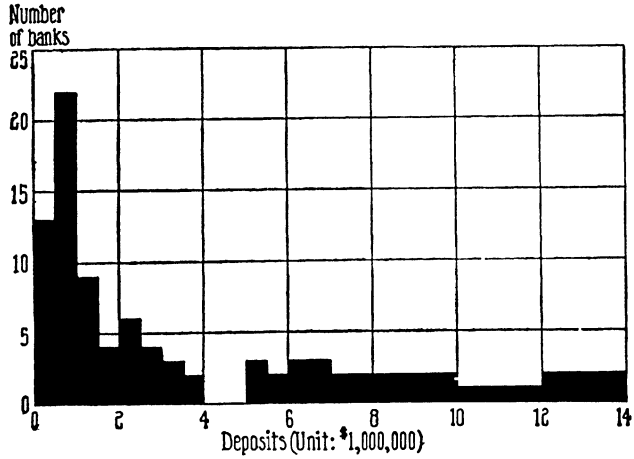


CHART 79.—Block diagram of the series of Chart 65.
(Data in Table *R*, Appendix A.)

The regrouping process[1] consists in combining the original frequencies in successively larger class intervals, with a view to removing the irregularities, and noting how the apparent mode shifts in the process (Table 73). The apparent mode moves from 9.00–9.49 to 9.00–9.99, to 8.50–9.49, to 9.00–10.49, to 8.50–9.99, to 8.00–9.49, to 7.50–9.49, to 9.00–10.99, as we pass to the right. The device is likely to be laborious, occasionally leads to a highly artificial result, and is not applicable in all cases. At best, it only discovers within which class interval the mode lies.

The mode is the most difficult to find of the elementary averages. It has an important place in statistical method because of its useful descriptive nature: it states the commonest value of the variable, and it is that value which is frequently most impor-

[1] See BOWLEY, "Elements of Statistics," p. 97.

TABLE 73

STUDY OF SHIFTING OF MODE BY REGROUPING FREQUENCIES IN DIFFERENT INTERVALS*

Wages*	f	Employees: 1-dollar intervals		Employees: 1½-dollar intervals			Employees: 2-dollar intervals	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Less than 2.00	0							
2.00– 2 49	0	0	0	0				0
2 50– 2.99	0				1		1	
3.00– 3 49	1	5	1		5			
3.50– 3.99	4			10				
4 00– 4.49	5		9			18		19
4 50– 4 99	9	14			29		33	
5 00– 5.49	15		24	33				
5 50– 5 99	9	24				46		
6.00– 6.49	22		31		61			76
6 50– 6.99	30	52		91			100	
7 00– 7.49	39		69			120		
7 50– 7.99	51	90			141			197
8 00– 8 49	51		102	158			231	
8.50– 8.99	56	107				180		
9 00– 9.49	73		129		177			
9 50– 9 99	48	121		169				224
10 00–10 49	48		96			151		
10 50–10 99	55	103			140		188	
11 00–11.49	37		92	134				
11 50–11.99	42	79				101		
12 00–12 49	22		64		77		83	114
12 50–12 99	13	35		41				
13.00–13 49	6		19			26		
13 50–13.99	7	13			16			
14 00–14 49	3		10	16				22
14 50–14 99	6	9				15	22	
15.00–15 49	6		12		15			
15 50–15 99	3	9						
16 00–16 49	4		7	13				
16 50–16.99	0	4				7		13
17 00–17 49	1		1	2	5		8	
17 50–17 99	1	2				3		
18 00–18 49	1		2		3			4
18 50–18.99	1	2		3			4	
19 00–19 49	1		2			2		
19.50–19 99	0	1			1			
20.00–20 49	0		0	1				2
20 50–20 99	1	1				2	2	
21.00–21 49	1		2		2			
21.50–21 99	0	1		1				1
Total	672	672	672	672	672	672	672	672

* Wages of weavers, in dollars per week, for the whole United States, as given in "Twelfth Census of the United States, 1900, Special Report on Employees and Wages," Washington, Bureau of the Census, 1903.

tant in the practical problems of applied statistics. Except in the form in which it appears in the complicated mathematical analysis of the curve-fitting problem, it is not readily available for algebraic manipulation. This defect, together with the lack of precision in its ordinary determination, must largely account for its limited use.

OTHER AVERAGES OF GROUPED FREQUENCY SERIES

The harmonic and geometric averages of frequency series can be computed by obvious adaptations of the process used in finding the mean (Tables 74 and 75). In both cases, class intervals which

TABLE 74
CALCULATION OF HARMONIC AVERAGE FOR THE SERIES OF TABLE 67

Weekly earnings X	Number of bookkeepers f	Center of class interval c	f/c
15–20	88	17 5	5.02857
20–25	229	22 5	10.17778
25–30	372	27 5	13 52727
30–35	139	32.5	4.27692
35–40	54	37.5	1.44000
40–45	14	42.5	0.32941
45–50	10	47.5	0.21053
50–55	15	52 5	0.28571
55–60	1	57 5	0.01739
60–65	2	62.5	0.03200
65–70	1	67 5	0.01481
70–75	1	72 5	0.01379
Total	926	. . .	35.35418

$$\text{Harmonic average } H = \frac{926}{35.35418} = 26.19 \text{ units}$$

were of uniform width for purposes of finding the mean are no longer uniform. This is true because the *variable* for the harmonic average is really the reciprocal of that for the mean, and the variable for the geometric average is really the logarithm of that used for finding the mean. These observations are significant also in connection with the assumption (which now must be slightly modified) that the variates within any particular class interval are distributed in such way that their mean is at the mid-point of that interval.

TABLE 75

CALCULATION OF THE GEOMETRIC AVERAGE FOR THE SERIES OF TABLE 67

Weekly earnings	Number of bookkeepers f	Center of class interval c	log c	$f \cdot \log c$
15–20	88	17 5	1 2430	109.3840
20–25	229	22.5	1 3522	309 6538
25–30	372	27 5	1 4393	535 4196
30–35	139	32 5	1 5119	210.1541
35–40	54	37 5	1 5740	84 9960
40–45	14	42.5	1 6284	22 7976
45–50	10	47 5	1 6767	16 7670
50–55	15	52 5	1 7202	25 8030
55–60	1	57.5	1 7597	1.7597
60–65	2	62.5	1 7959	3.5918
65–70	1	67 5	1.8293	1.8293
70–75	1	72 5	1 8603	1 8603
Total	926	....	.....	1324 0162

$$\log G = \frac{1324.0162}{926} = 1.4298. \quad G = 26.90 \text{ units}$$

TABLE 76

CALCULATION OF THE EXPONENTIAL AND INVERSE EXPONENTIAL AVERAGES FOR THE SERIES OF TABLE 67

Weekly earnings X	Number of bookkeepers f	c	$c/100 = x$	e^x	$f \cdot e^x$	e^{-x}	$f \cdot e^{-x}$
15–20	88	17 5	0 175	1 1913	104 8344	0 8395	73 8760
20–25	229	22 5	0 225	1 2523	286 7767	0 7985	182 8565
25–30	372	27 5	0 275	1 3165	489 7380	0 7596	282 5712
30–35	139	32 5	0 325	1 3841	192 3899	0 7225	100 4275
35–40	54	37 5	0 375	1 4550	78.5700	0 6873	37.1142
40–45	14	42 5	0 425	1 5296	21.4144	0 6538	9.1532
45–50	10	47 5	0 475	1 6080	16 0800	0 6219	6 2190
50–55	15	52.5	0 525	1 6904	25.3560	0 5915	8 8725
55–60	1	57.5	0 575	1 7771	1 7771	0 5627	.5627
60–65	2	62 5	0.625	1 8683	3.7366	0.5352	1 0704
65–70	1	67 5	0 675	1 9640	1 9640	0 5092	5092
70–75	1	72 5	0 725	2 0648	2.0648	0 4843	.4843
Total	926				1224.8019		703.7167

Exponential average = E, where $e^E = 1224.8019/926 = 1.3227$
$E = 0.2797$ units, or 27.97 dollars
Inverse-exponential = I, where $e^{-I} = 703.7167/926 = 0.7600$
$I = 0.2028$ units, or 20.28 dollars

Many other types of average could be derived by setting up appropriate definitions.[1] Specialized types of averages serve in the analysis of series of particular kinds. For general statistical analysis, only the elementary averages are needed, and of these the mean and median and mode are of chief importance.

---

[1] Table 76 shows two of these, the *exponential* and the *inverse-exponential averages*, for which the definitions are implied in the formulas

$$e^E = \frac{\Sigma f e^x}{N} \quad \text{and} \quad e^{-I} = \frac{\Sigma f e^{-x}}{N}$$

respectively, wherein E is the exponential average and I is the inverse-exponential average; e, the Napierian logarithmic base (2.7128 . . .); c, the center of the class intervals; and x, a measure derived from c by the application of a factor k (= 0.01, in this case) so chosen that $e^k = a$, where a is the required "base" for the average It may be possible to choose another constant h so that e^{kx+h} will fall within the scope of the available exponential table; or upon certain necessary considerations any available logarithm table may be used instead of the exponential table, or to supplement the exponential table when such table is too "short."

CHAPTER XII

DISPERSION

MEASURES OF DISPERSION DERIVED DIRECTLY FROM THE ARRAY

It was remarked in Chap. IX that the characteristic of a frequency series, next in importance after the average, is the dispersion. *Dispersion* is that property of a series by which the several variates tend to differ in value from the average. Two series having identical means but different dispersions appear in Chart 80. The object of measuring dispersion is to secure a single summary number which adequately indicates the extent of scatter of the variates—presumably *all* of the variates, though in some appraisals of dispersion not all variates are considered—from the average. A summary number of this sort is essentially an average and should therefore have the six basic requisites of a good average.

The crudest measure of dispersion is the range, the difference between the largest and smallest variates. The fact that the extreme variates are generally more susceptible to irregular influences than the intermediate variates renders the range peculiarly erratic. Moreover, because the range depends only upon the two extreme variates of the series and takes no account of the scatter of the other variates, it is not a satisfactory summary number.

The force of this objection can largely be removed by taking the range between some other pair of variates than the two extremes. A common practice is to use one-half the interquartile range. This measure is called the *quartile deviation*, Q. The *quartiles* (see Chap. IX) are values of the variable located on a principle similar to that which determines the median. Just as the median divides the total frequency in halves, the *lower quartile* Q_1, divides the frequency in the ratio 1:3, and the upper quartile, Q_3, divides the frequency in the ratio 3:1. The evaluation of the quartiles rests upon steps parallel to those taken in getting the median. Q_1 must be located so that, if the variates are in an array, one-quarter of them fall below it; and likewise Q_3 must be located so that three-quarters of the arrayed variates fall below it.

The first step in determining the quartiles therefore consists in calculating $n/4$ and $3n/4$, where n is the total frequency.

In actual practice, some ambiguity attaches to the location of the quartiles; but this ambiguity is not practically serious except when the number of variates, n, is small. The ambiguity arises because, when n is a multiple of 4, each quartile falls between—and

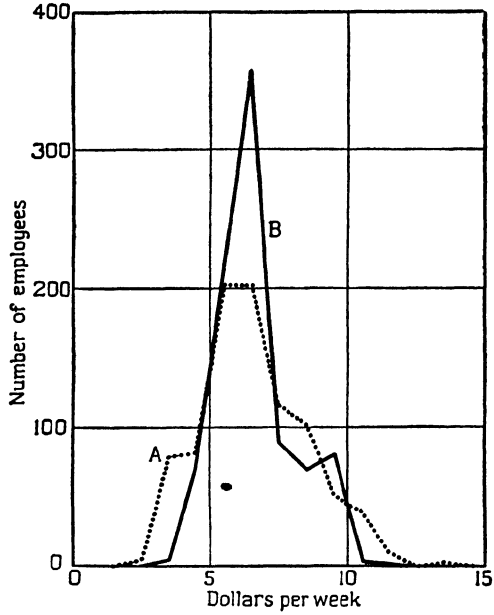


CHART 80.—Distribution of wages of female dresser tenders and beamers (*A*) and females in all other occupations (*B*), in woolen mills in 1900—in the United States for *A*, in New England states for *B*.

at an uncertain point between—two variates of the array. Suppose a series consists of 12 variates in an array; $n/4$ is 3; 3 variates fall below Q_1 and 9 fall above it; and Q_1 lies *somewhere* between the third and fourth variates of the array.

The ambiguity disappears in any case where n is not a multiple of 4—where $n/4$ and $3n/4$ are not whole numbers. Of course, the fractional part does not strictly fall "below" Q_1; but, if we placed Q_1 above the indicated variate, we should put too much frequency below Q_1. Suppose, for example, there are 11 variates in an array; $n/4$ is $2\frac{3}{4}$. As $2\frac{3}{4}$ variates must fall below Q_1, and $8\frac{1}{4}$ above Q_1, we conclude the third variate in the array is Q_1. Sup-

TABLE 77

NET PREMIUMS WRITTEN BY 75 STOCK MULTIPLE LINE INSURANCE COMPANIES FOR 1937*

Company	Net premium
The Travelers Insurance Company	63,552[a]
Hartford Accident & Indemnity Company	38,462
United States Fidelity & Guaranty Company	35,011
Aetna Casualty & Surety Company	31,291
Employers' Liability Assur. Corp., Ltd.	28,891
Maryland Casualty Company	27,926
Fidelity & Casualty Company of New York	27,412
Continental Casualty Company	20,690
General Accident Fire & Life Assur. Corp. Ltd.	19,950
Aetna Life Insurance Company	17,798[a]
Globe Indemnity Company	16,713
Standard Accident Insurance Company	16,445
Zurich General Accident & Liab. Ins. Company	15,612
Massachusetts Bonding & Insurance Company	14,644
New Amsterdam Casualty Company	14,332
Royal Indemnity Company	13,821
American Auto. Insurance Company	13,307
Indemnity Insurance Company of N. A.	12,472
Travelers Indemnity Company	*11,277*
Fidelity & Deposit Company of Maryland	10,826
Great American Indemnity Company	10,305
Ocean Accident & Guarantee Corp. Ltd.	9,945
European General Reins. Company, Ltd.	9,660
American Surety Company	9,535
National Surety Corporation	9,340
Employers Reinsurance Corp.	8,526
London Guarantee & Accident Co. Ltd.	8,467
Hartford Steam Boiler Inspection & Ins. Co.	7,306
Century Indemnity Company	7,190
Glens Falls Indemnity Company	6,875
Commercial Casualty Company	6,732
Pennsylvania Mfrs. Assn. Cas. Ins. Company	6,467
United States Casualty Company	6,422
American Motorists Insurance Company	6,307
American Employers Insurance Company	5,945
Metropolitan Casualty Insurance Company	5,933
Fireman's Fund Indemnity Company	5,794
General Reinsurance Corp., New York	*5,716*
United States Guarantee Company, New York	5,484
Ohio Casualty Insurance Company	5,482

* Unit: thousand dollars. Source: "1938 Argus Chart," Cincinnati, The National Underwriter Company, 1938, p. 169. "Multiple Line" means that two or more of the following types of business have been written: Liability, Auto. Liability, Workmen's Compensation, Fidelity, Surety, Plate Glass, Burglary and Theft, Accident and Health, Property Damage and Collision, Auto. Property Damage, Auto. Collision, Steam Boiler, Machinery, Credit, Water Damage, Fire and Tornado, etc.

Median and quartile items are in italics.

[a] Excludes premiums of life department.

TABLE 77 (*Continued*)

Company	Net premium
Pacific Indemnity Company	5,423
Saint Paul-Mercury Indemnity Company	5,376
Bituminous Casualty Corp.	5,216
Pacific Employers Insurance Company	4 944
Associated Indemnity Corp.	4,573
Preferred Accident Insurance Company	4,397
Central Surety & Insurance Corp.	4,153
Bankers Indemnity Insurance Company	4,021
Standard Surety & Casualty Company of New York	3,906
Trinity Universal Insurance Company	3,895
Phoenix Indemnity Company	3,878
London & Lancashire Indem Company of America	3,869
New Jersey Mfrs. Casualty Insurance Company	3,849
Western Casualty & Surety Company	3,577
Eagle Indemnity Company	3,410
Manufacturers' Casualty Insurance Company	3,324
Sun Indemnity Company of New York	*3,321*
National Casualty Company	3,275
American Re-Insurance Company	3,104
National Auto. Insurance Company	3,042
American Fidelity & Casualty Company, Va.	3,018
Commercial Standard Insurance Company	2,901
Home Indemnity Company	2,871
New York Casualty Company	2,664
General Casualty Company of America	2,581
American Casualty Company	2,511
Employers Casualty Company	2,451
Car and General Insurance Company	2,437
Columbia Casualty Company	2,320
Allstate Insurance Company	2,167
Traders & General Insurance Company	2,154
American Credit Indem Company	2,073
American States Insurance Company	2,056
Buckeye Union Casualty Company	1,916
Keystone Auto. Club Casualty Company	1,905

pose further that n is 10; $n/4$ is $2\frac{1}{2}$, and $2\frac{1}{2}$ variates fall below Q_1 while $7\frac{1}{2}$ fall above it. Hence, the third variate of the array is Q_1. Suppose finally that n is 9; $n/4$ is $2\frac{1}{4}$, and $2\frac{1}{4}$ variates fall below Q_1 while $6\frac{3}{4}$ fall above it. Hence, the third variate is Q_1. In these cases there is no doubt about identifying Q_1 with a particular variate. As already remarked, the ambiguity, where it exists at all, is not serious if n is fairly large. (Note the italicized specification of the quartiles for Table 77.)

The spacing assumption.—The foregoing considerations satisfactorily determine the quartiles for a series which is truly discrete —for a series pertaining to a discontinuous variable. The situation becomes more complicated if the variable is continuous, as is strictly the case for Table 77; and considerations pertinent to a "discrete" series which does not record a truly discontinuous variable will provide the basis for a method of estimating the quartiles for a grouped frequency series described below.

Consider the upper quartile of Table 77. As n is 75, $3n/4$ is 56¼: thus 56¼ variates fall "below" the upper quartile. If the variable were truly discontinuous, the variate for the Travelers Indemnity Company would be regarded as the upper quartile. The variable in this case is, however, not discontinuous—the observations might have taken on any values whatever (waiving the limitation of expression in indivisible money units). We therefore make a basic assumption: that the observation for the Travelers Indemnity Company pertains, not to a single point of 11,277 thousand dollars, but to a *range* of the variable extending from a point midway between 10,826 and 11,277 to a point midway between 11,277 and 12,472. That is: the range from 11,051½ to 11,874½ is assumed to belong to, or be "covered" by, the 57th variate. That the 57th variate happens to fall at a specific point, 11,277, within that range is from now on ignored. We observe, incidentally, that the actual variate, 11,277, does not fall at the midpoint, 11,463, of the assumed range; and this disparity is also apparent in the case of the lower quartile, for which the actual variate (3,321) differs from the midpoint (3,310¼) of the range between 3,298 and 3,322½. But these disparities are ignored—we make the flat assumption that the single variate applies to the range reaching from a point midway between it and the next smaller variate to a point midway between it and the next higher variate.[1]

We can now estimate the upper quartile on the basis of this assumption. Below 11,051½ we assume that 56 variates fall, and identify the space between 11,051½ and 11,874½ with the 57th variate. The upper quartile position therefore falls ¼ of the way from 11,051½ to 11,874½—that is, at 11,257¼. As such great

[1] This assumption is, of course, an approximation; but, particularly near the middle of the entire range of variation where the quartiles are likely to occur, the error involved is not likely to be large. If the same procedure were used for a position near the end of the entire range, such as the first or ninth decile, the error might be serious, particularly if the total frequency n were not large.

precision is obviously not warranted, this result would be rounded to 11,257, and perhaps to 11,260.

In similar manner, the lower quartile has 18¾ variates falling below it. Our assumption places 18 variates below 3,298, the midpoint between 3,275 and 3,321. Therefore the lower quartile position falls ¾ of the way from 3,298 to 3,322½—that is, at 3,316⅜. And this would be rounded to 3,316, or perhaps to 3,320.

According to our spacing assumption the lower and upper quartiles are not 3,321 and 11,277, as indicated in Table 77, but 3,316 and 11,257.[1]

Quartiles of a grouped frequency series.—The spacing assumption, with a slight modification, lies at the basis of the estimate of a quartile for a grouped frequency series. Here, as in the case of the median (page 183), we assume the frequency in a particular class interval—the interval within which the quartile has been located by examining the schedule of cumulative frequencies—is uniformly distributed within that interval. In other words, if the frequency in the interval is k, we assume that the interval is divided into k equal subintervals, and that one of the k variates belongs to each of these subintervals.[2]

In locating the quartile within the particular class interval, the procedure is fairly obvious. The cumulative frequency for all classes below that interval is assumed to belong to the range of the variable exactly reaching to the lower boundary of that interval. Within that interval, each of the k variates is supposed to cover a range reaching over the single subinterval to which that variate is assigned. Thus, to find the lower quartile of the series in Table 78 we identify Q_1 as the $92\frac{6}{4}$th variate—that is, the 231.5th variate. There are 88 variates—according to the cumulative series $\underline{F}$—falling below earnings of $20, and reaching, on the spacing assumption, exactly up to $20; and Q_1 falls somewhere in the 20–25 interval. In fact Q_1 is the 143.5th variate in that interval; and, on the spacing assumption which assigns the 229 variates of that interval to 229 subintervals of equal length, this quartile position is

---

[1] We may now reconsider the location of the median, as carried out in Chap. XI. If the same spacing assumption were applied in connection with Table 72, the median would appear as 26.97 instead of 26.96 as there given.

[2] At either end of the class interval this assumption may not be—probably is not—exactly equivalent to the assumption above for the categorical series of Table 77. The point is that if the frequencies in the two adjacent intervals (assumed of equal length) are not equal, the last subinterval in one class interval will not be exactly as long as the first subinterval in the next class interval. This error is, however, not likely to be serious in any frequency series having substantial frequencies in all intervals near the quartiles.

143.5/229 of the length (5 dollars) of the class interval above the lower boundary (20 dollars). Hence Q_1 is 20 + 3.13 = 23.13.

By similar analysis, we could reckon Q_3; and either can be reckoned by using the downward cumulative series $\bar{F}$ instead of the upward series $\underline{F}$.

TABLE 78

CALCULATION OF THE QUARTILE DEVIATION FOR THE SERIES OF TABLE 67

Weekly earnings X	Ordinary frequency f	Cumulative frequency $\underline{F}$	Cumulative frequency $\bar{F}$
15–20	88	88	926
20–25	229	317	838
25–30	372	689	609
30–35	139	828	237
35–40	54	882	98
40–45	14	896	44
45–50	10	906	30
50–55	15	921	20
55–60	1	922	5
60–65	2	924	4
65–70	1	925	2
70–75	1	926	1
	926		

$Q_1 = {}^{926}/_{4}$th item; the 231.5th variate in the $\underline{F}$ column: or the 694.5th variate in the $\bar{F}$ column. Since there are 88 variates below 20 and 609 variates from 25 up, Q_1 is determined by the value of the 143.5th variate from 20 toward 25 or of the 85.5th variate from 25 toward 20. Now, assuming that each of the 229 variates within this interval "covers" a subinterval of length $^5/_{229}$, Q_1 is found to be \$23 13, whether calculations are made from 20 toward 25 or from 25 toward 20. Thus

$Q_1 = 20 + {}^5/_{229}143.5 = 20 + 5 \times .6266 = 23.13$; or $= 25 - {}^5/_{229}85.5 = 23.13$.

Similarly Q_3, the median, is \$26.97. And Q_3, which is the 694.5th variate, is the 5.5th variate in the 30–35 class, and therefore

$$Q_3 = 30 + {}^5/_{139}5.5 = 30.20.$$

With Q_1 and Q_3 determined the quartile deviation is found by the formula

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

and is, for Table 78, 3.54.

A weakness of all these *position* measures, based upon the range or portions thereof, is that the precise size of most of the variates has no effect on the result. For example, the quartile deviation

will be the same whether the variates between Q_1 and Q_3 are concentrated just above Q_1 or are spread uniformly from Q_1 to Q_3, so long as Q_1 and Q_3 are unchanged. The fault is clearly important in the measurement of a characteristic of the series, especially that characteristic which relates directly to the scattering of the variates from the central type (average). Two measures designed to meet this objection are the average deviation and the standard deviation.

THE AVERAGE DEVIATION

The *average deviation* (*A. D.*) is, strictly, the mean of absolute (without regard to plus and minus signs) deviations of the several variates from the median. In actual practice this strict definition is usually modified, in that deviations are measured from the mean rather than from the median, because of the somewhat greater facility of measuring deviations from the mean. The computation consists in obtaining the deviation of every variate from the median (or mean), dropping all minus signs, and calculating the mean of these deviations.[1]

---

[1] The algebraic formula can be developed on general lines, and then applied to each case. Suppose (Chart 81) that the vertical line represents the variable scale,

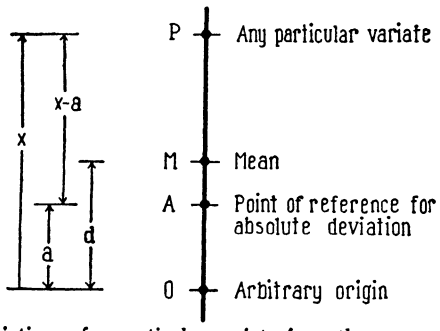


CHART 81.—Deviations of a particular variate from the mean and from certain other points.

that A is the point from which absolute deviations are to be measured, that O is an arbitrary origin of measurement. Let a be the distance from O to A, d the distance from O to M, and x the value of any variate as measured from O (all measurements in some arbitrary unit). For any particular variate the deviation from A is $x - a$; and this is negative for all variates falling below A, and positive for all falling above A. (This is equally true whether A lies above or below O.) The negative character of the deviations for variates below A is removed by expressing them as $a - x$. Hence the average deviation from A is

THE STANDARD DEVIATION

The labor of calculating the average deviation is considerable, and the several steps are often confusing. The chief reason, however, for preferring the standard deviation is that it is better adapted to algebraic manipulation and has a more direct relation to the theoretical analysis of frequency distributions than the

---

$$\frac{1}{N}\left\{\sum_{x\leqslant a}(a-x)f_x+\sum_{x>a}(x-a)f_x\right\}$$

where f_x is the frequency of the variate x, N is the total frequency, and $\sum_{x\leqslant a}$ indicates "sum of all terms like . . . , for which x is less than or equal to a."

The numerator of the above expression transforms to

$$a\sum_{x\leqslant a}f_x-\sum_{x\leqslant a}xf_x+\sum_{x>a}xf_x-a\sum_{x>a}f_x$$

and, if N^* represents the total frequency equal to or less in value than A, that is, if

$$N^*=\sum_{x\leqslant a}f_x$$

and if we subtract and add $\sum_{x\leqslant a}xf_x$ this becomes

$$aN^*-2\sum_{x\leqslant a}xf_x+\sum_{x\leqslant a}xf_x+\sum_{x>a}xf_x-a(N-N^*)$$

which is

$$a(2N^*-N)-2\sum_{x\leqslant a}xf_x+Nd$$

since

$$\sum_{x\leqslant a}xf_x+\sum_{x>a}xf_x=\Sigma xf_x=Nd$$

by definition of mean. Hence the fraction, found above for the average deviation from A, is

$$a\frac{2N^*-N}{N}-\frac{2}{N}\sum_{x\leqslant a}xf_x+d$$

In practice a is known as soon as A is selected, and d is a number ordinarily computed in the analysis of every series. Hence the second term must be evaluated: the values of x for all variates at or below A are added together, and this total is multiplied by 2 and divided by N. If N^* is not known by inspection, it must be found by cumulating all the frequencies up to and including that at A. For a grouped frequency series, this means that if A has a numerical value equal to or larger than the class mark of a class interval the frequency in that interval is counted, and otherwise not. In other words, for this computation, it is assumed that all the variates within an interval lie exactly at the center of that interval. This assumption of exact "centering" is clearly much more stringent than the distribution assumptions made for finding the averages and quartiles.

For the average deviation *from the median* the term $\frac{a2N^*-Na}{N}$ drops out (not precisely, if N is odd, for then this term becomes a/N) and the expression for

average deviation. The *standard deviation* is the square root of the mean of the squares of the deviations from the mean.

---

the average deviation simplifies to

$$A.D._c = -\frac{2}{N}\sum_{x<a} xf_x + d$$

For the average deviation *from the mean*, a becomes d, and the first and third terms combine to $2N^*d/N$, giving

$$A.D._M = \frac{2}{N}\left(N^*d - \sum_{x<a} xf_x\right)$$

It is this second form, the average deviation from the mean, which is the more commonly used in practice. On theoretical grounds, however, the first form is the more logical; for, as was shown in Chap. XI, the sum (and therefore the mean) of the absolute deviations from the median is less than from any other point. Table 79 shows the calculations for the series of Table 67. (The foregoing discussion follows that given in Kelley's text (*op. cit.*, p. 71), which should be consulted for the details.)

TABLE 79

CALCULATION OF THE AVERAGE DEVIATION FOR THE SERIES OF TABLE 67

Weekly earnings X	Number of bookkeepers f	x	$f \cdot x$	
			−	+
15–20	88	−2	−176	
20–25	229	−1	−229	
25–30	372	0		
30–35	139	1		139
35–40	54	2		108
40–45	14	3		42
45–50	10	4		40
50–55	15	5		75
55–60	1	6		6
60–65	2	7		14
65–70	1	8		8
70–75	1	9		9
Total	926		−405	441

Average deviation about the mean, $A.D._M = 2/N\left(N^*d - \sum_{x'<a} xf_x\right)$

$d = {}^{36}\!/_{926}$, $N^* = 317$, and $\sum_{x'<a} xf_x = -405$

Hence, $A.D._M = {}^{3}\!/_{926}({}^{1141}\!/_{926} + 405) = 0.9013$ class intervals $= 4.507$

While the average deviation about the median, $A.D._c = -\frac{2}{N}\sum_{x'<a} xf_x + d$

The median = \$26.96, as in Table 72

Hence, $A.D._c = {}^{405}\!/_{463} + 0.0389 = 0.9136$ class intervals = \$4.568

$A.D._c$ is here greater than $A.D._M$ because data are in grouped frequency form (see page 185).

TABLE 80

CALCULATION OF THE STANDARD DEVIATION OF A CATEGORICAL SERIES BY DIRECT APPLICATION OF THE DEFINITION*

Serial number	Loans X	$X - M_a$		$(X - M_a)^2$
		−	+	
1	549	85		7,225
2	1584		950	902,500
3	1938		1304	1,700,416
4	1786		1152	1,327,104
5	950		316	99,856
6	305	329		108,241
7	602	32		1,024
8	100	534		285,156
9	210	424		179,776
10	227	407		165,649
11	165	469		219,961
12	495	139		19,321
15	232	402		161,604
16	64	570		324,900
17	107	527		277,729
18	1019		385	148,225
21	127	507		257,049
22	120	514		264,196
23	94	540		291,600
24	418	216		46,656
25	220	414		171,396
26	1492		858	736,164
27	2653		2019	4,076,361
28	2131		1497	2,241,009
29	1742		1108	1,227,664
30	862		228	51,984
31	64	570		324,900
32	240	394		155,236
33	282	352		123,904
34	690		56	3,136
35	438	196		38,416
36	397	237		56,169
37	1772		1138	1,294,044
38	1497		863	744,769
39	326	308		94,864
40	428	206		42,436
41	190	444		197,136
42	1658		1024	1,048,576
43	3874		3240	10,497,600
44	105	529		279,841

TABLE 80 (*Continued*)

Serial number	Loans X	$X - M_s$		$(X - M_s)^2$
		−	+	
1	137	497		247,009
2	249	385		148,225
3	5181		4547	20,675,209
4	150	484		234,256
5	334	300		90,000
6	130	504		254,016
7	86	548		300,304
8	189	445		198,025
9	756		122	14,884
10	176	458		209,764
1	60	574		329,476
2	207	427		182,329
3	33	601		361,201
4	1069		435	189,225
5	631	3		9
6	97	537		288,369
7	1021		387	149,769
8	288	346		119,716
9	48	586		343,396
10	796		162	26,244
11	86	548		300,304
12	269	365		133,225
13	232	402		161,604
14	370	264		69,696
15	493	141		19,881
16	99	535		286,225
17	165	469		219,961
18	946		312	97,344
19	1043		409	167,281
20	223	411		168,921
21	221	413		170,569
22	201	433		187,489
23	875		241	58,081
24	94	540		291,600
25	120	514		264,196
26	18	616		379,456
27	136	498		248,004
28	574	60		3,600
29	131	503		253,009
Total	50,087	22,752	22,753	58,036,665

Mean equals $\frac{50087}{79} = 634.01$

Standard deviation $= \sqrt{\frac{58,036,665}{79}} = \sqrt{734,641} = 857.1$

* Unit for loans: thousand dollars. Data pertain to loans and discounts of Michigan banks at end of 1936, as explained in Table 61.

A direct application (Table 80) of this definition will yield the result; but the labor, for any except the simplest series, is very extensive. In calculating the standard deviation of a grouped frequency series (Table 81), all the variates in any one class interval are assumed to lie exactly at the center of the interval—the same stringent assumption as used in getting the average deviation.

TABLE 81

CALCULATION OF THE STANDARD DEVIATION FOR THE DATA OF TABLE 67

Weekly earnings X	Number of bookkeepers f	c	fc	$c - M_x$	$(c - M_x)^2$	$f(c - M_x)^2$
15–20	88	17 5	1,540 0	−10 194	103 9176	9,144 7488
20–25	229	22 5	5,152 5	− 5 194	26 9776	6,177 8704
25–30	372	27 5	10,230 0	− 0 194	0 0376	13 9872
30–35	139	32 5	4,517 5	4 806	23 0976	3,210 5664
35–40	54	37 5	2,025 0	9 806	96 1576	5,192 5104
40–45	14	42 5	595 0	14 806	219 2176	3,069 0464
45–50	10	47 5	475 0	19 806	392 2776	3,922 7760
50–55	15	52 5	787 5	24 806	615 3376	9,230 0640
55–60	1	57 5	57 5	29 806	888 3976	888 3976
60–65	2	62 5	125 0	34 806	1,211 4576	2,422 9152
65–70	1	67 5	67 5	39 806	1,584 5176	1,584 5176
70–75	1	72 5	72 5	44 806	2,007 5776	2,007 5776
Total	926		25,645 0			46,864 9776

$$\text{Mean} = \frac{25{,}645\ 0}{926} = 27.694 \text{ units}$$

$$\text{Standard deviation, } \sigma = \sqrt{\frac{46{,}864.9776}{926}} = 7.114 \text{ units}$$

The labor of deriving the standard deviation can be reduced greatly by using the *skeleton method.* As in the calculation of the mean, an arbitrary origin is selected at the center of a centrally located class interval, and a new unit is chosen, equal to the width of one class interval. The deviations x from the arbitrary origin constitute, with the frequencies f_x, an artificial series of which the standard deviation σ (*sigma*) is found by the following formula[1]:

[1] By definition,

$$\sigma = \sqrt{\frac{\Sigma x^2 f_x}{N}}$$

where x is a deviation from the mean. It was shown in Chap. XI that, since

$$\sigma = \sqrt{\frac{\Sigma x^2 f_x}{N} - d^2}$$

This formula can be applied much more easily than the original definition, because it is unnecessary to square numerous decimal fractions. When σ has been evaluated in terms of the arbitrary unit, it is converted to the original unit by use of the proper multiplier (see Table 82).

TABLE 82

RECALCULATION, BY THE SKELETON METHOD, OF THE STANDARD DEVIATION ALREADY DERIVED IN TABLE 81

Weekly earnings X	Number of bookkeepers f	x	fx		x^2	fx^2
			−	+		
15–20	88	−2	−176		4	352
20–25	229	−1	−229		1	229
25–30	372	0			0	0
30–35	139	1		139	1	139
35–40	54	2		108	4	216
40–45	14	3		42	9	126
45–50	10	4		40	16	160
50–55	15	5		75	25	375
55–60	1	6		6	36	36
60–65	2	7		14	49	98
65–70	1	8		8	64	64
70–75	1	9		9	81	81
Total	926		−405	441		1,876

$d = {}^{8}\%_{926} = 0.0389$ class intervals, or 0.1945 units. Hence $M = 27.69$ units

Standard deviation, $\sigma = \sqrt{187\%_{926} - (0.0389)^2} = 1.4228$ class intervals, or 7.114 units

COEFFICIENTS OF DISPERSION

The three principal measures of dispersion discussed above are the quartile deviation, the average deviation, and the standard deviation. Each of these measures has the same dimension as the original variates: if those variates are expressed in terms of length

$$x = z + d$$
$$\Sigma x^2 f_x = \Sigma z^2 f_z + d^2 N$$

from which

$$\sigma = \sqrt{\frac{\Sigma z^2 f_z}{N} - d^2}$$

(or money value or weight or age), the dispersion measures are also in terms of length (or money value or weight or age, as the case may be). Although this fact is not objectionable in the description of a single series, it is a serious obstacle to the comparison of several series. Moreover, even if two series of variates involve the same type of dimension, the *units* of measurement may be different, or the sizes may be on a distinctly higher level in one series than in the other. In all these cases, nevertheless, some direct comparison of dispersion is desirable—and such comparison should be independent of these outside influences.

The *abstract* specification of dispersion, independent of the dimension or unit or general level of the variates, may be effected by citing numbers known as *coefficients of dispersion.* A coefficient in this sense is an abstract number.[1] The usual device available for eliminating the dimension from a measure is division by an appropriate magnitude of similar dimension. Coefficients of dispersion are generally obtained by dividing a measure of dispersion by the appropriate average.

Thus, Q is divided by the mean of Q_1 and Q_3 to yield the *quartile coefficient*

$$\frac{Q}{\frac{Q_1 + Q_3}{2}} = \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_1 + Q_3}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

The *A.D.* should logically be divided by the median on the ground that the median is the point from which deviations are measured in getting the *A.D.* As, however, such deviations are in practice measured from the mean, the mean is taken as the divisor. The standard deviation is divided by the mean, to yield the *coefficient of variation*

$$\frac{\sigma}{M}, \text{ or } 100\frac{\sigma}{M}\%$$

The last coefficient is the most common, and it is usually expressed in percentage form. Coefficients afford significant comparisons of dispersions in many instances where direct measures would be misleading. One of two series may have a larger coefficient of variation than the second, whereas the second may have a larger standard deviation than the first.

[1] Compare this concept with the definition of coefficient as used in describing the terms of an algebraic expression. The essential point is that the coefficient is a mere number quite independent of the nature of the variate.

CHAPTER XIII

THE NORMAL LAW OF ERROR

THE EQUATION AND PROPERTIES OF THE NORMAL CURVE

Common experience shows that several measurements of the same physical magnitude are in general not all equal. The causes of variation among the results are likely to be numerous: the effects of temperature and atmospheric conditions upon the measuring instrument differ for the different observations; the angle from which the observer views the instrument may vary; lost motion in the adjustment devices of the instrument may be taken up with varying success; the judgment of the observer may vary as he estimates fractions upon his scales; and many other influences may work to produce fluctuations in the readings.

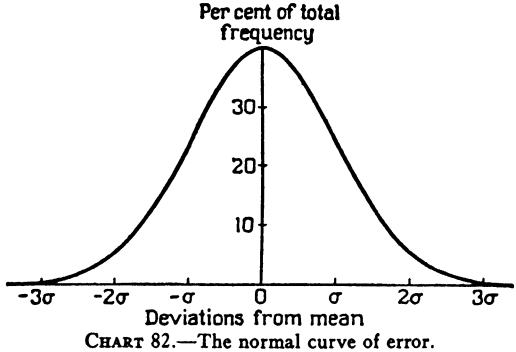


CHART 82.—The normal curve of error.

In general these disturbing influences have two characteristic properties: each is as likely to increase as to decrease the observed size, and each is more likely to produce small than large effects upon the result. If a large number of measurements of a given magnitude are made under the conditions outlined, and if a frequency series is formed with these measurements as variates, the corresponding frequency curve tends to take the form shown in Chart 82. As the differences among the several measurements

are assumed to result from "errors" of observation, the standard frequency curve of Chart 82 is called the *normal error curve.*[1]

Certain important properties of the curve appear from its equation. The largest value of y is that for which x is zero, and therefore the mode of the curve coincides with the mean. Since $-x$ yields the same y as $+x$, the curve is symmetrical about the vertical line through $x = 0$. Therefore half the frequency lies on each side of that vertical line, and this implies that the median coincides with the mean. There is no finite value of x for which y is 0, but y is exceedingly small for all values of x outside of $\pm 3\sigma$ (see Table 145 in which values of $y\sigma/N$ are given for various values of x/σ). It can be shown (by methods involving the integral calculus) that one-half the total area between the curve and the x-axis lies between two vertical lines distant $\pm .67449\sigma$ from O (see Table 145 for values of the area corresponding to various values of x/σ). In view of the area concept (Chap. IX) of frequency, this statement is seen to imply that quartiles are at a distance $\pm .67449\sigma$ from O and that, therefore, $Q = .67449\sigma$ for this curve. It can be shown (also by operations involving calculus) that the average deviation for this series is, approximately,

$$A.D. = 0.7979\sigma$$

By extension of these findings, the approximate statements can be made, for a unimodal frequency curve which is *nearly* normal, that Q is about $\frac{2}{3}\sigma$, and that $A.D.$ is about $\frac{4}{5}\sigma$.

FITTING A NORMAL CURVE TO A GIVEN SERIES

The equation of the normal law clearly implies that the form of the curve depends only upon σ, assuming N fixed. In other words, series having the same frequency but different dispersions will have correspondingly different normal curves (see Chart 131). Therefore σ is called the *parameter* of the normal curve: to each value of σ belongs a particular normal curve from the whole family

---

[1] It is also called the *Gaussian error curve* (in honor of the mathematician Gauss), the *normal probability curve*, and the *curve of the normal law of error*, and frequently also the *normal curve*. It is in such a curve that the relation between the abscissa (deviation from the mean) and the ordinate (frequency) for any point is

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

where N is the total frequency, σ the standard deviation, e the Napierian constant 2.71828, π the circular constant 3.14159, x the abscissa, and y the ordinate.

of possible curves of this type. The curves of this family differ from each other only with respect to their dispersions: the dispersion is the single distinctive fact about a particular normal curve, as compared to other normal curves.

As soon as σ is known, a normal curve can be "fitted" to a frequency series. This assumes, of course, that such fitting is appropriate, and the essential requisite is that fluctuations among the variates have the same nature as errors of observation, although the particular series often has nothing to do with a group of measurements of a single physical magnitude. Usually the

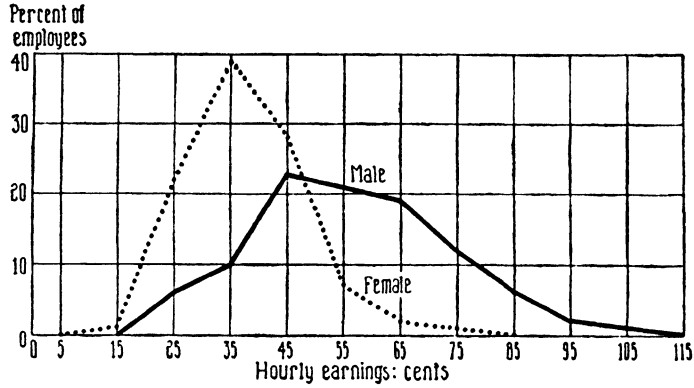


CHART 83 —Polygons comparing relative frequencies of male and female employees—hand and machine treers—in the boot and shoe industry, according to hourly earnings in 1930.

(Data in Table 42, page 124.)

preliminary test consists in examining the chart of the given series to ascertain whether it is substantially of normal form. For example, the "female" curve of Chart 83 (reproducing Chart 44) appears as a fairly good approximation to the normal type, whereas the "male" curve is not.

The essential computations comprise finding the mean and standard deviation of the series; and these computations, for the female series of Chart 83, are carried out in Table 83.

Knowledge of the mean is essential in order to locate the normal curve in the right position along the horizontal axis: the mode—which is also the mean—of the normal curve which fits the given series must lie at the mean, $0.378, of the given series. Knowledge of the standard deviation is essential in order to specify the parameter, $\sigma = 1.04$ class intervals, of the particular normal curve which fits.

Substitution of the results of Table 83 in the general formula for the normal curve gives

$$y = \frac{100}{1.04\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{1.04}\right)^2}$$

as the equation of the fitted curve. Here x is the distance from the mean to the particular value of the variable under study, measured in class intervals. Thus

$$x = \frac{X - M_x}{10}$$

To "fit" the normal curve to the given series, the above equation is used to secure y for each particular X. This computation

TABLE 83

CALCULATION OF MEAN AND STANDARD DEVIATION FOR THE "FEMALE" SERIES OF CHART 83 AND TABLE 42*

Hourly earnings X	Per cent of employees f	x	fx		x^2	fx^2
			−	+		
15	1	−2	− 2		4	4
25	22	−1	−22		1	22
35	39	0				
45	28	1		28	1	28
55	7	2		14	4	28
65	2	3		6	9	18
75	1	4		4	16	16
Total	100		−24	52		116
			+28			

$d = {}^{28}/_{100} = 0.28$ class intervals, or 2.8 cents: $M = 35 + 2.8 = 37.8$ cents per hour

$\sigma = \sqrt{{}^{116}/_{100} - (0.28)^2} = \sqrt{1.16 - 0.0784} = 1.04$ class intervals, or 10.4 cents

* Earnings entries are in dollars, for center of each class interval

can be carried out by logarithms, but much time and effort are saved by using compiled tables of the normal function, such as column b of Table 145. In this table, u corresponds to x/σ in the general equation of footnote on page 209, and the items in column b are not for y of that equation but for $y\sigma/N$. One procedure, from this point on, consists in taking from column b the items corresponding to selected stubs in column a—say, for $u = 0, 0.5, 1,$

1.5, 2, 2.5, 3, and 3.5. These items multiplied by N/σ, which is 100/1.04 or 96.2 in the present problem, are the values of y—the ordinates of the fitted normal curve—corresponding to the selected values of u. Table 84 shows in column 2 the items taken from column b of Table 145, and in column 3 the corresponding y. Columns 4 and 5 enable us to determine the corresponding points on the horizontal axis: column 4 is obtained by multiplying column 1 by σ (in cents—that is, 10.4), and column 5 follows at once. Points corresponding to (X, y) can then be plotted on the chart.

TABLE 84

CALCULATION OF NORMAL ORDINATES FOR THE SERIES OF TABLE 83, AT POINTS X SYMMETRICALLY SPACED FROM THE MEAN

u	$\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$	y	$X - M_x$	X
(1)	(2)	(3)	(4)	(5)
−3 5	0 001	0 1	−36 40	1 4
−3 0	0 004	0 4	−31 20	6 6
−2 5	0 018	1 7	−26 00	11 8
−2 0	0 054	5 2	−20 80	17 0
−1 5	0 130	12 5	−15 60	22 2
−1 0	0 242	23 3	−10 40	27 4
−0 5	0 352	33 8	− 5 20	32 6
0	0 399	38 4	0.00	37.8
0 5	0 352	33 8	5 20	43 0
1 0	0 242	23 3	10 40	48 2
1 5	0 130	12 5	15 60	53 4
2 0	0 054	5 2	20 80	58 6
2 5	0 018	1 7	26 00	63 8
3 0	0 004	0 4	31 20	69 0
3 5	0 001	0 1	36 40	74 2

The several points are next joined by a smooth curve, and the result can be compared with the polygon for the actual data.

Comparable ordinates.—The foregoing procedure, while it takes the material from Table 145 in the simplest and easiest way, does not (unless M happens to fall at the center of a class interval of the given series) yield values of y corresponding to—and hence directly comparable with—the given frequencies f. The difficulty is that the points on the horizontal axis to which the various y belong are not centers of the given class intervals. To overcome this difficulty, the u for which items of column b—and from them the corresponding y—are obtained from Table 145 must be chosen so that these u belong to centers of given class intervals.

Accordingly, the particular u, instead of being chosen arbitrarily as 0, 0.5, 1, etc., are *calculated* as follows:

$$u = \frac{x}{\sigma}$$

and x measured in class intervals, where

$$x = \frac{X - M_x}{10}$$

X being the value of the variable corresponding to the center of a class interval. More briefly, if σ is *now* stated in the original units of the variable ($\sigma = 10.4$),

$$u = \frac{X - M_x}{\sigma}$$

Computations on this basis, for the series of Table 83, appear in Table 85. Column 3 gives the u corresponding to the centers of

TABLE 85

CALCULATION OF NORMAL ORDINATES FOR THE SERIES OF TABLE 83, AT CENTERS OF CLASS INTERVALS

X	$X - M_x$	$\frac{x - M_x}{\sigma} = u$	$\frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$	y	f
(1)	(2)	(3)	(4)	(5)	(6)
5 00	−32 8	−3 15	0 0028	0 3	0
15 00	−22 8	−2 19	0 0363	3.5	1
25 00	−12 8	−1 23	0 1872	18 0	22
35 00	− 2 8	− 27	0.3847	37.0	39
37 80	*0 0*	*0 00*	*0.399*	*38 4*	—
45 00	7 2	0 69	0 3144	30 2	28
55 00	17 2	1 65	0 1023	9.8	7
65 00	27 2	2 62	0 0129	1 2	2
75 00	37 2	3 58	0 0007	0.1	1

the given class intervals, and column 4 gives the corresponding items taken from column b of Table 145. Column 5 gives the corresponding y: each item of column 4 is multiplied by N/σ, which is 96.2. (Note especially that to get N/σ, one uses σ in class intervals, *not* in cents.) Column 6 repeats the given frequencies from Table 83—they are now correctly comparable with the ordinates of the fitted curve in column 5.

Either of the foregoing computations yields the same fitted normal curve. It is shown, along with the polygon of original data, in Chart 84 (reproduced from Chart 48). As the two types of computation have been carried out, that of Table 84 yields more points of the curve and therefore a more precise plotting. We could, however, introduce more details in Table 85—for example, by taking X for the end of each class interval as well as for the center. Of course, the method admits of finding y for *any* X we please, and for as many values of X as we choose. When the method of Table 85 is used, the ordinate should always be cal-

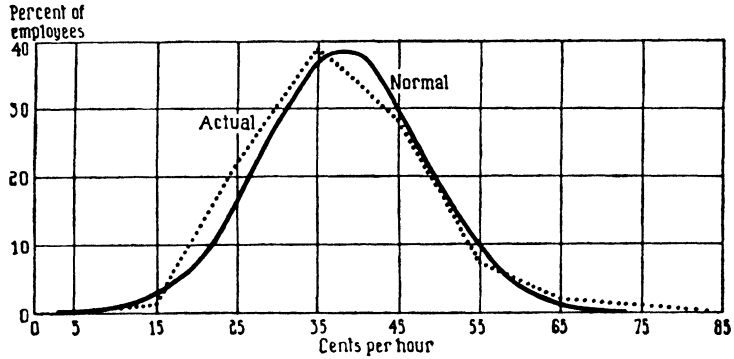


CHART 84.—Polygon of actual data, and fitted normal curve, for "female" series of Chart 44.

(Data in Table 43, page 126.)

culated for x equal to zero—by the method of Table 84—because that is the ordinate at the mode, the *maximum* ordinate.

An important advantage of Table 85 is that it yields values of y properly comparable with the given f: it shows how the given f would need to be changed in order to fit exactly the normal curve. The differences between y and f (columns 5 and 6 of Table 85) measure the discrepancies between the actual frequencies and corresponding frequencies of the fitted normal curve. An entirely different method of calculating the normal-curve frequencies—*not* ordinates—belonging to the given class intervals is the *method of areas*. It rests upon column a of Table 145, but is not explained and illustrated here.

The determination of the normal curve belonging to a particular series is an instance of *smoothing* by means of a mathematical formula. If the observed group of cases covered by the actual data is regarded as a sample from a larger population for which

complete data are not available, the normal curve—assuming that it fits well—may be taken as the frequency curve of that larger population. The deviations—discrepancies—of the polygon of actual data from the normal curve are then regarded as *errors of sampling;* they are the irregularities in the given series due to the inadequacy of the sample upon which it is based. The fitted normal curve, or its equation, may be regarded as a device for calculating the theoretical frequencies which the sample would have if it were truly representative, or to interpolate frequencies for intermediate values of the variate (in a discrete series) or for narrower class intervals (in a continuous series).

STANDARD UNITS

The unit in which x is measured is identical with that in terms of which σ is expressed: if σ is expressed in class intervals, x should be measured in class intervals. It is often desirable to express x (as measured from the mean) in terms of one standard deviation as unit, and the variable is then said to be expressed in *standard units.* Thus any series of variates, X, may be expressed in standard units by first subtracting the mean of the series from each X and dividing the resulting differences by the standard deviation of the series. Thus if u represents the measurement X expressed in standard units, $u = \dfrac{X - M}{\sigma}$ where M and σ, as usual, represent the mean and the standard deviation of the series.

MEASURES OF DISPERSION, AND THE PROBABLE ERROR

The relations between Q, σ, and $A.D.$ for the normal curve are of wide practical import. For the normal curve, Q is approximately $\frac{2}{3}\sigma$, and $A.D.$ is approximately $\frac{4}{5}\sigma$. As these relations hold approximately for unimodal curves differing moderately in form from the normal type, they furnish a basis for checking results or for estimating one measure from another, for such series. On the other hand, they furnish an indirect test of the closeness of conformity of a given series with the normal type; for, with proper reservations, the more nearly these relations are realized, the more closely does the series follow the normal form.

Of chief importance, however, is the practical use made of the Q: σ relation in the applications of theory of sampling. In this connection Q is generally called the *probable error, p.e.;* and this concept, taken over from the theory of errors of observation,

occupies a prominent place in all practical statistical analysis. By definition

$$p.e. = Q = .67449\sigma$$

Moreover, since, by definition of the quartiles, half the total frequency lies between the quartiles—in the interval between $x = \pm p.e.$—if any one variate is selected at random from the N given variates there is a fifty-fifty chance that the selected case falls within a range extending $.67449\sigma$ on each side of the mean.

Using the terminology of errors, this proposition takes the following form: There is a fifty-fifty chance that the "error," in the sense of deviation from the mean, of a variate chosen at random is less than the *p.e.* This statement assumes that the frequency distribution of variates is normal; and it does not apply to distributions differing markedly—because of skewness, bimodality, or other reasons—from the normal.

SELECTION OF A RANDOM SAMPLE

Suppose that a given series of actual data is a random sample from a larger group, which may be called the *population.* A *random sample* from a population of individual objects (or instances) is a sample formed in such way that every one of the individuals in the population has the same chance of being selected in the sample and that the selection of a particular individual does not influence the chance of selecting some other particular individual. In practice such sample is usually chosen by chance; but, as the manner in which the element of chance enters may vary, not all selections by chance are truly random samples. In contrast to the random sample is the *sample by design,* in which the selection is made upon some arbitrary plan designed to yield a sample which is representative of the population. Although sampling by design is used widely in practical statistics, it is only to *random* sampling that the doctrine of chance—the theory of error—can properly be applied. It is only for a random sample that the theory of error yields a numerical estimate of the probable size of error in the statistical characteristics of the population as inferred from the sample. Another term sometimes applied to a sample is *representative;* but no generally accepted precise definition is available, and it is often used as a "weasel word" by those who have no critical knowledge of the sampling methods used.

One of the practical means of selecting a sample at random is by drawing lots. The serial numbers up to a certain point are

written on slips or cards and shuffled in a box, and a selection is made from them by drawing. Table 86 gives a sample of 100 cases chosen from the numbers 1 to 400 by this method. To obtain a random sample of 100 cases from a population of 400 given individuals, the 400 individuals are arranged in an entirely hap-

TABLE 86

SAMPLE OF 100 NUMBERS DRAWN BY LOT FROM THE SERIAL NUMBERS 1–400, AND ARRANGED IN ORDER OF SIZE

1	97	170	241	312
6	100	174	246	316
13	108	177	252	320
14	111	181	254	323
16	115	183	255	326
22	118	189	259	335
29	120	191	262	338
31	124	196	265	343
34	127	203	272	347
42	132	205	277	350
46	137	208	278	356
49	139	214	280	357
51	140	219	284	361
59	145	221	287	364
64	153	223	293	367
76	158	224	298	373
80	160	229	301	379
84	162	231	304	381
88	168	233	307	385
93	169	235	310	398

hazard order, serial numbers 1 to 400 are assigned, and those cases having numbers given in Table 86 may then be taken as constituting the sample. It is particularly important that the original arrangement of the 400 cases be haphazard, so that the serial order of the cases shall not impart a bias to the sample.[1]

The frequency curve of the sample generally differs notably from that for the entire group. Chart 85 compares the curve for a sample of 100 employees with that for the entire group of 400

[1] The selected sample is not truly random, in any case; for the total frequency of the entire group is not very large compared with the frequency of the sample. Hence the selection of the earlier individuals of the sample influences the chance that some other individual will be selected. Thus the chance that an individual *A* will be chosen is 1:400 at the start, but becomes 1:301 just before the last member of the sample is drawn. This situation could have been avoided if each serial number drawn had been put back into the box before drawing the next. Very likely, in that case, some numbers would have been drawn more than once. Such multiple inclusions of a case from a small population in the random sample is wholly consistent with the chance notion upon which random sampling rests.

employees, the frequencies having been changed proportionately so that the total is the same for both curves (see page 123). In general, the characteristics of the sample differ from similar characteristics for the whole group. Thus the mean of the sample is 6.34 and that of the entire group (population) is 6.03.

This discrepancy between the characteristic of a sample and the corresponding characteristic of the whole group is the central object of study in the theory of sampling. The point of practical importance is that the data for the whole group are seldom available (though they happened to be so for this illustration), and the true characteristic is never known. The data of the sample must therefore be taken as representative of the whole group, and the characteristic derived from the sample must be accepted as the approximate value of the unknown characteristic of the whole group. Obviously, some estimate of the error involved in this approximation is desirable, and the theory of sampling aims to afford such estimate.

PROBABLE ERRORS OF THE CHARACTERISTICS

Suppose the conditions of Chart 85 are modified so that the total number of employees in the population becomes very large, that not one but many random samples are taken (though the total number of employees included in the aggregate in all the samples need not be equal to the total of the whole population: the samples, in other words, do not form a comprehensive and mutually exclusive set of selections from the population). If, then, the mean of each sample is found, and if these means are classified in a frequency series (having *its* total frequency equal to the number of samples), such frequency series will be nearly of the normal type.[1] This series will have a mean, the mean of the "means of the samples"; and this mean will be a very close approximation to the unknown mean of the whole group (population). In fact, if the number of samples is large, this mean is so close to the unknown mean of the whole group that the error is negligible, and in practice the two may be assumed identical.

---

[1] The fact that the fundamental frequency distribution (of wages for all employees in the entire group, in this case) is not normal, or even very nearly normal, does not alter this rule. The frequency distribution of the means of the several samples will nevertheless be very nearly normal, provided certain requirements which are by no means stringent are met by the fundamental frequency distribution and by the process of sampling (see F. Y. EDGEWORTH, "The Law of Error," *Cambridge Philosophical Transactions*, Vol. 20).

The frequency series of means has also a standard deviation, and this standard deviation, σ_M, is subject to the following approximate relation

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

and therefore the probable error of the mean is

$$p.e.(M) = 0.67449\frac{\sigma}{\sqrt{N}}$$

where σ is the standard deviation of a sample, and N the number of cases in the sample.[1]

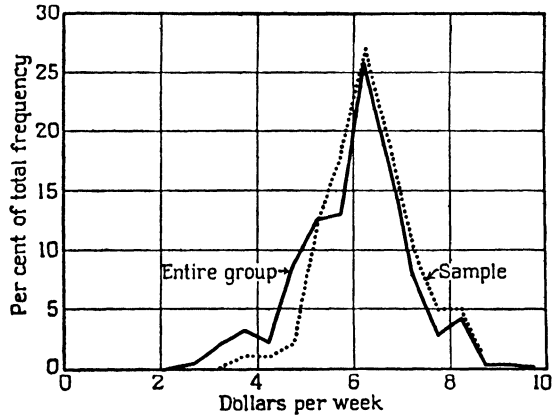


CHART 85.—Comparison of a sample distribution with the distribution of the general population from which the same was drawn.
(Data in Table T, Appendix A.)

As the distribution of the means is approximately normal, it is accurate to say that there is a fifty-fifty chance that the mean of any sample lies within $0.67449\frac{\sigma}{\sqrt{N}}$ of the mean of the entire group from which the sample was drawn. This gives some estimate of the possible error in assuming M, as computed from a single sample, to be the mean of the entire group.

Just as a frequency distribution of the means of the several samples was made, one can establish a frequency distribution of

[1] Strictly, σ is the standard deviation for the fundamental distribution of the entire population; but as that measure is unknown, the value of σ for the single sample is assumed in its stead, without serious danger of large error.

values of a higher characteristic; and, by similar reasoning, show that there is a fifty-fifty chance that the value of that characteristic for any one sample lies within one *p.e.* of the unknown value of that characteristic for the entire group. By analysis parallel to that which showed the *p.e.* of the mean approximately equal to $0.67449\frac{\sigma}{\sqrt{N}}$, the *p.e.* of the standard deviation is found to be approximately $0.67449\frac{\sigma}{\sqrt{2N}}$.

The normal law of error enters in arriving at the above conclusions, in that the frequency series of the means of samples is taken as of normal type. This condition is essential to the statement that the mean of a particular sample has a fifty-fifty chance of falling within one *p.e.* of the general mean. Now, an important empirical fact is that this second condition is closely fulfilled even for phenomena which are not essentially normal: even if the frequency series of actual data deviates quite widely from the normal type, the frequency series of the means (or values of certain other characteristics) of the samples tends to be nearly normal (see footnote on page 218). Upon this consideration rests the justification for applying the sampling theory—including its algebraic and numerical aspects—to statistical analyses of frequency series which differ quite markedly from the normal type.

CHAPTER XIV

SKEWNESS, MOMENTS

COEFFICIENTS OF SKEWNESS

The next important characteristic of a frequency series, after the dispersion, is the skewness. *Skewness* is that property of a series by which the variates tend to be dispersed more on one side of the mean than on the other. Chart 84 compares a moderately skew (asymmetrical) frequency curve with a normal curve having the same mean and the same standard deviation. The mode of the skew curve is to the left of the mode, or mean (the mode of the

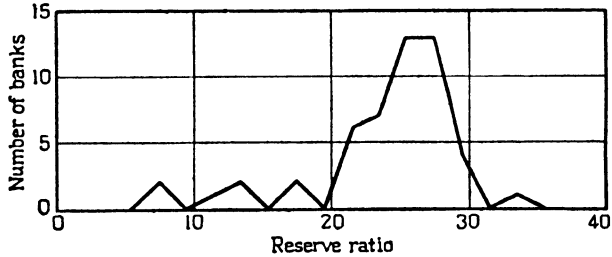


CHART 86.—Reserve ratios of New York City Clearing-House banks for the week ending October 26, 1907.

(Reproduced from page 213 of Crum and Patton's "Economic Statistics," 1925 ed.)

normal curve is always coincident with the mean), of the standard symmetrical curve. Of the extremities of the skew curve, the lower is closer to the mean. Moreover, the median of the skew curve is at the left of the mean, the median being 36.92 whereas the mean is 37.80; and this is equivalent to saying that less than half the total frequency lies above the mean of the skew curve. A skew curve distorted in the direction shown in Chart 84 is *negatively* skew, and one in which the distortion is in the inverse direction (Chart 86) is *positively skew*.[1] The skew curve of Chart 84 has moderate skewness, whereas that of Chart 86 has large skewness.

This relative distortion on the two sides of the mode may vary in extent, and a measure (or rather, a coefficient) of this distortion

[1] Some authors, for example, Karl Pearson and W. P. Elderton, assign these terms in the opposite sense.

from the normal curve is called the *skewness*. For unimodal curves which are only moderately skew, an approximate value for the skewness, Sk, can be obtained by one of the following formulas:

$$Sk = -\frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

$$Sk = -\frac{mean - mode}{\sigma}$$

$$= -3\frac{mean - median}{\sigma}$$

In each case the numerator is an expression which tends to be larger according as the distortion is greater, and the denominator is a measure of dispersion.

A fourth, and more precise, formula, which however involves more laborious calculations, is

$$Sk = -\frac{1}{\sigma}\sqrt[3]{\frac{\Sigma f_x x^3}{N}}$$

where x is measured from the mean. This is the algebraic definition of skewness, and is somewhat analogous to that of dispersion, except that the index of the root and of the power of x is 3 instead of 2. This formula may be expressed in words as, "The negative of the ratio of the cube root, of the mean of the cubes of the deviations of the variates from their mean, to the dispersion." The labor of direct calculation from the formula is almost prohibitive, but the skeleton method may be used, since the above formula is algebraically identical with

$$Sk = -\frac{1}{\sigma}\sqrt[3]{\frac{\Sigma f_x x^3}{N} - 3d\sigma^2 - d^3}$$

where x is measured from an arbitrary origin and where d is the distance of the mean from the arbitrary origin. The corresponding calculations appear in Table 87, and the results based upon the approximate formulas are appended below the table.

The formulas given above for Sk are *coefficients* of skewness. Each compares a measure of the amount of distortion (lack of symmetry), as given by its numerator, with a measure of the dispersion, as given by its denominator. The reason for preferring the coefficient to a direct measure of the mere amount of distortion

is that two curves similar in shape would have different *amounts* of skewness (the numerator of any of the above formulas) if their dispersions were not equal. Their *coefficients* of skewness should nevertheless be equal.

TABLE 87
CALCULATION OF SKEWNESS FOR THE SERIES OF TABLE 67

Weekly earnings X	Number of bookkeepers f	x	fx		fx^2	fx^3	
			−	+		−	+
15–20	88	−2	176		352	704	
20–25	229	−1	229		229	229	
25–30	372	0					
30–35	139	1		139	139		139
35–40	54	2		108	216		432
40–45	14	3		42	126		378
45–50	10	4		40	160		640
50–55	15	5		75	375		1875
55–60	1	6		6	36		216
60–65	2	7		14	98		686
65–70	1	8		8	64		512
70–75	1	9		9	81		729
Total	926		−405	441	1876	−933	5607
			36			4674	

$d = 0.0389$ class intervals; $\sigma = 1\ 4228$ class intervals, or 7.114 units

$$Sk = -\frac{1}{1.4228}\sqrt[3]{\frac{4674}{926} - 3(0.0389)(1.4228)^2 - (0.0389)^3} = -1.186$$

$Q_1 = 23.12$, $Q_2 = 26.96$ = Median, $Q_3 = 30.18$

Mean = 27.69, Mode = 27 (about)

Hence,

$$-\frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} = 0.0878; \qquad -\frac{\text{Mean} - \text{Mode}}{\sigma} = -0.0970;$$

$$-3\frac{\text{Mean} - \text{Median}}{\sigma} = -0.3078$$

EXTREMELY SKEW SERIES

The first three formulas above are good estimates of the skewness—as strictly defined by the fourth formula—only if the frequency distribution is but moderately skew. These approximate formulas apply properly to series which are fairly similar to the normal curve. Although some economic variables—such as the wages case covered in Table 87—yield frequency series having only moderate skewness, many frequency series derived from economic data are extremely skew.

No satisfactory general definition of "moderate," as applied to degree of skewness, can be given in precise terms. Coefficients of skewness, calculated by the fourth formula above, might be compared with some standard—say a *Sk* of 1 or 1.5—and cases having *Sk* less than such arbitrarily chosen number could be called *moderately skew.* But such a procedure is scarcely satisfactory because, as shown below, a few extremely large items can yield a

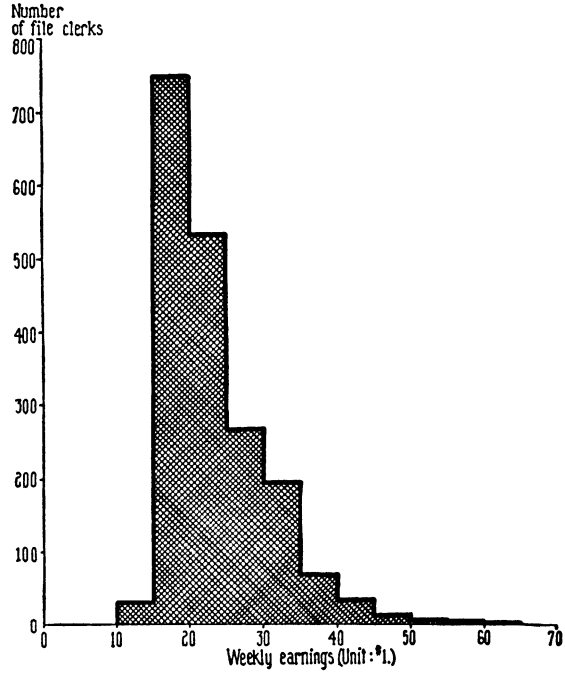


CHART 87.—Block diagram of weekly earnings of file clerks in New York City in May, 1937.

(Data in Table 41, page 122.)

large *Sk* whereas the general sweep of the curve throughout the main range of variation is only moderately skew. A tolerably satisfactory basis of appraisal is the chart of the given frequency series. Thus, as already noted, the curve of Chart 84 is moderately skew and that of Chart 86 is more than moderately skew.

Chart 87 shows another moderately skew distribution, again from the field of wage statistics. The group of employees for which this series is given is fairly homogeneous—the group includes workers in a narrowly defined occupation, in a single community,

at a single time. Even wage statistics, for a nonhomogeneous group of workers, are likely to yield extremely skew distributions. Thus, if we regard as "wages" the entire labor income of each resident of the United States, the corresponding frequency curve will surely be extremely skew because of the exceptionally high salaries (including commissions, fees, and bonuses) of a small fraction of the population. Precise information for constructing the entire frequency distribution for this case is not available, but we can get approximately accurate data for those individuals filing Federal tax returns.[1] This fragment of the series is extremely skew; and as the wage earners who do not file tax returns presumably have labor incomes generally much smaller than the tax-exemption limit, we may be very certain that the frequency distribution for the entire population, if it were available, would be even more extremely skew.

As already remarked, a tolerably satisfactory basis of determining whether a frequency distribution is only moderately skew is study of the chart of the data; and the statistician will, by experience in examining many charts, secure a sense of what is moderate skewness. A graphic device which sometimes aids in studying the skewness of a particular series, though it does not indicate whether such skewness is only moderate, consists in plotting the horizontal measurements—variates—on a logarithmic chart. Chart 88 is an example of such plotting. The effect is to reduce the *apparent* skewness: the farther a point is to the right on an arithmetic scale, the greater its shift toward the left on the log scale. If the skewness were precisely of the sort—as is approximately true for some economic variables—for which the resulting curve on the log scale is normal, we would conclude that although the variates themselves have a skew distribution the logarithms of the variates have a normal distribution. For such a series, the geometric average is unmistakably preferable to the mean, as a typical summary number.

A defect of Chart 88 is that the horizontal spaces between plotted points are wider at the left than at the right. This flows from the fact that the given class intervals were of equal width, and the logarithms of the centers of class intervals therefore become closer and closer as we pass to the right. This defect would disappear if the class limits of the original tabulation, instead of being

[1] See, for example, column 1 of table on p. 13 of "Statistics of Income, 1934, Part 2," Washington, U. S. Treasury, 1936.

equally spaced, were spaced so that the differences between the logarithms of successive limits were equal. Sometimes frequency data for skew series are in fact tabulated and published for such intervals, or intervals nearly following this rule.

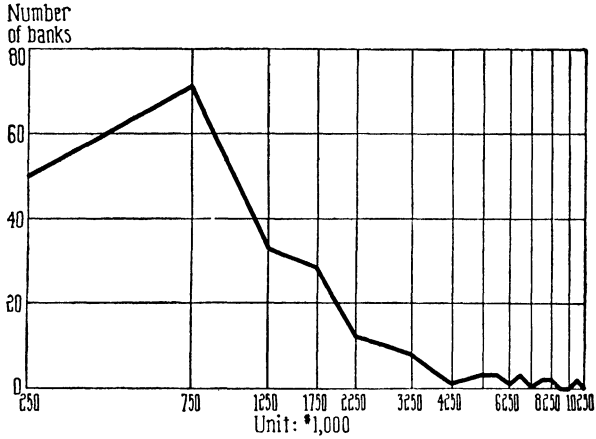


CHART 88.—Frequency polygon for distribution of Ohio national banks according to deposits, on December 31, 1936
(Data in Table Q, Appendix A.)

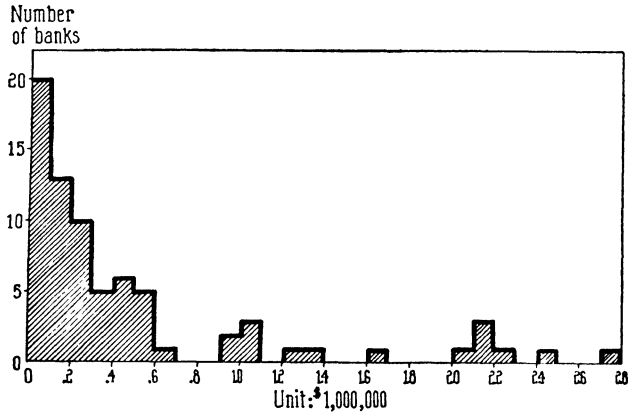


CHART 89.—Number of Michigan national banks having holdings of U. S. Government securities in stated amounts on December 31, 1936.
(Data in Table *U*, Appendix A.)

Economic variables often reflect a type of variation which can only result in a very skew series; and the statistician should expect his economic frequency series to be very skew and perhaps

extremely skew, rather than nearly normal. Such large skewness generally appears in series directly or indirectly reflecting differences in size of economic units, such as the size of the particular enterprises in an industry. Not infrequently the skewness takes such an extreme form that the series has no mode at all—its maximum frequency occurs, not within the range of variation, but at one end thereof. Such a series is called *J-shaped;* an illustration appears in Chart 89. In this particular case, the J-shapedness may be apparent rather than real: if the \$0–\$100,000 interval were broken into subintervals, their frequencies might rise for a brief range just right of 0 and thus indicate the existence of a mode. The contrary might, however, be the case: the series might still appear J-shaped, even for the narrower classification.

MOMENTS OF A FREQUENCY SERIES

The mean, standard deviation, and coefficient of skewness are examples of characteristics based upon the moments of the series. The notion of moment in statistics is analogous to that of mechanics: the first *moment*, about any point O, of a particular frequency, f_x, is the product of that frequency expressed as a fraction of the total frequency (somewhat analogous to force in mechanics) by the corresponding deviation (arm) from the point O. Thus the moment for the frequency in the interval having the deviation x from O is

$$x\frac{f_x}{N}$$

The sum of these products, for the entire series, is called the *first* moment—or moment of the first *order*—of the series about the point O. This first moment is designated by μ_1 (mu), and is

$$\mu_1 = \frac{1}{N}\sum xf_x$$

If the frequency is multiplied by the second, rather than the first, power of the deviation, the result is the *second moment* about O

$$\mu_2 = \frac{1}{N}\sum x^2f_x$$

Similarly, if the cube of the deviation is used as a factor, the *third moment* about O is obtained

$$\mu_3 = \frac{1}{N}\sum x^3 f_x$$

And, in general, a moment of any order may be defined as the sum of the products of f_x/N by the appropriate power of x.

These moments about any chosen point O are called *crude moments*, and are those actually derived in computation. For description of the series, either as bases of definitions of characteristics or as parameters in the equations of frequency curves, moments about the mean—rather than about any point O—are needed; these are called *principal moments*, and are represented by the symbol ν (nu). Thus

$$\nu_1 = \frac{1}{N}\sum \bar{x} f_x$$

$$\nu_2 = \frac{1}{N}\sum \bar{x}^2 f_x$$

$$\nu_3 = \frac{1}{N}\sum \bar{x}^3 f_x$$

The definitions of the first three characteristics—mean, standard deviation, and coefficient of skewness—give therefore

$$\mu_1 = d, \qquad \nu_1 = 0$$
$$\nu_2 = \sigma^2, \qquad \nu_3 = -\sigma^3 Sk^3$$

As previously remarked in the case of ν_2 and ν_3 (pages 206 and 222), each of the principal moments can be expressed in terms of the corresponding crude moment and principal moments of lower order. This is a property due to the symmetry of the expansion of powers of a binomial, because $\bar{x}$ is $x - d$. One great practical advantage of this property is that characteristics, defined in terms of principal moments, can be computed from crude moments which are derivable by the skeleton method. A further advantage comes in the extensive algebraic analysis of the frequency series of theoretical statistics, where, in reducing results to a form convenient for computation, principal moments often need to be replaced by crude moments.

CURVE FITTING BY THE USE OF MOMENTS

The notion of moment serves as a guide in elaborating a theory of curve fitting for skew curves. As was remarked earlier (Chap. IX), one of the requisites in smoothing a frequency curve is that the total area under the smoothed curve equal that of the original

block diagram; this is another way of saying that the total frequency should be unchanged. As the total frequency may be looked upon as a moment of zero order, this is equivalent to requiring that the zero moment of the smoothed curve equal that of the actual series. Likewise, we may require that moments of higher orders be identical for the two curves. This affords a means of determining the parameters of the mathematical equation of the smoothed curve in terms of the moments of the actual frequency series. The successive determination of those moments yields constants for the equations of curves which successively fit the series more closely. Thus it has been shown that ν_2 (or σ^2) determines the normal curve which most nearly fits the given skew series. The further determination of ν_3 (involving Sk) yields a smooth skew curve, which fits better than the normal curve.

Formally this process may be carried on indefinitely by getting moments of higher and higher order. In practice, however, it is seldom that any improvement in the fit is obtained by using moments above the sixth, and in most cases the analysis should not proceed beyond the fourth moment. A limitation upon the practical use of the process is imposed by the rapid increase in the probable error of the moment, as the order of the moment increases. Even for the first moment (mean), an extremely large or small value of the variate has undue influence on the result. For the second moment this extreme variate has a much greater relative effect because of the squaring process (note the squares of large items in Table 80): the extreme deviation is "weighted" by a large number, that deviation itself, whereas a moderate deviation is weighted by a small number, the moderate deviation. As moments of higher order involve raising the deviations to still higher powers, a single extreme variate may evidently dominate the values of the higher moments, and quite outweigh the other $N - 1$ variates. In some series such extreme variates are not present, and more of the moments may properly be used; but, even for series which appear free from irregularities, danger exists that the values of the higher moments will be erratic. Most series encountered in practice include some extreme variates; and although there may be no single very exceptional variate, moments above the fourth are likely to be quite unreliable.

CHAPTER XV

THE CORRELATION TABLE

DEFINITION OF CORRELATION

That some sort of interrelation exists between many of the simplest natural phenomena is well known, and we have a somewhat less confident opinion that economic and other social phenomena are linked together in groups. For example, some relation between the variations of rainfall, temperature, and crop production is generally accepted. Our first inclination in such cases is to assume that the relation is one of cause and effect; but, even in the instance cited, question may fairly be raised as to which is cause and which is effect. Does a variation in temperature depend on the rainfall? Is a variation in agricultural production due to a change in temperature, or should it be said that the general state of vegetation influences the temperature? In this particular example, the nature of the chief causal relation is quite clear; but the suggestion that, even in so simple a case, an inverse causal relation may hold is sufficient to emphasize our general ignorance. We are usually at a loss to determine the exact nature of the causes and effects involved in concomitant variations. The interest of the statistician is fixed less upon the question whether one phenomenon causes another than upon the discovery of a mutual relation between the phenomena and upon the measurement of the degree of this relation. Such a mutual relation between the variable phenomena is called *correlation.*

For example, suppose that it is asked whether correlation exists between the rate of earnings of common stocks and the price of those stocks. If generally the stock of a company having large earnings per share sells at a high price and that of a company having small earnings per share sells at a low price, correlation exists between price and earnings per share, and such correlation is *positive.* If, on the contrary, stock of companies having large earnings per share generally sell at a low price, and conversely, the correlation between price and earnings per share is *negative.* If there is no clearly evident tendency for large earnings per share to be associated with high prices, and conversely, there is no correla-

tion, or such correlation as may exist is so slight that its presence cannot be demonstrated.

Except in the rarest cases the existence of correlation does not imply a simple relation of proportionality between the two variable magnitudes. Thus if the earnings per share for one company is twice that for another company, the price of the stock of the first company is not likely to be exactly twice that for the second company. The point is that correlation in the vast majority of practical problems involves only an *approximate* dependence; whereas proportionality is a *precise* relation. On the other hand, when a well-defined correlation exists, it affords a basis for estimating one variable magnitude (say stock price of a particular company) from the other variable magnitude (say earnings per share of that company). The result is, however, only an estimate, and not a calculation of the same precision as one of strict proportionality.

The above illustration brings out certain elementary ideas concerning correlation. The two phenomena between which the correlation is sought (earnings and price) are called the *variables*, and a pair of values of the two variables, belonging to one object or instance (company), is called a pair of *associated* variates. Correlation is essentially an average of the relation between the associated variates, for all the pairs of values. Such relation, in its simplest form, is a ratio between the two variables, each measured from an appropriately selected origin. If this average is highly typical—if it is highly representative of the individual relation of every pair—correlation is said to be *high*, and otherwise, *low*. If, instead of two variables, there were three or more variables involved, the relation of any one of the variables to the others would involve multiple correlation. This more complicated topic will be reserved for later discussion (Chap. XVII).

DESCRIPTION OF THE CORRELATION TABLE AND SCATTER DIAGRAM

The discovery of the existence and estimation of the degree of correlation rest upon a study of the correlation table; and even when more precise methods of measuring correlation are to be used, the table should always be examined and taken as guide. A *correlation table* is a frequency table of the second order in which the stubs indicate classification according to the size of one variable and the captions according to the size of the other. Strictly, which variable runs horizontally (and which runs vertically) in the table is a matter of indifference, although we customarily find it con-

TABLE 88

PRICE AND EARNINGS PER COMMON SHARE, FOR SELECTED MACHINERY AND EQUIPMENT COMPANIES, 1937*

Serial number	Estimated earnings	Price	Serial number	Estimated earnings	Price
2	0 9	7 5	43	1.25	14
3	0 65	7	45	2	10
4	2 3	24	47	5	32
6	0 4	3	48[a]	4	68
7	0 3	3 5	49[ab]	2.57	15
8	2 2	43	50[a]	6	26
9	0 8	5 5	52	4 25	47
10	2 5	14	54[ab]	0 53	5
11	2 75	21	55[a]	4.3	25
12[c]	4 0	53	56[a]	8.	66
13	2	15	59[ab]	5.32	46
15	3.25	17	60	7	30
16	1.25	15	61	2 75	22
17[a]	2 75	14	62	1 75	17
18	1 75	6.5	63	3 5	13
19	4.25	27	64[c]	5 75	68
20	1	8	65	2 3	17
21	3	20	66	0 75	5.5
22	3 25	17	67[a]	2 75	13
23	5 5	51	70	0 6	4.5
24	2.5	11	71	6.5	54
25	3	7	72	4	43
26	1	7.5	73[a]	1	26
27	1.5	10	75	3 1	33
28	1.5	9	76	0.4	8.5
29	4	28	77[ab]	1 75	9
30[ab]	4.11	30	78	4.5	42
32	3	14	79	2	23
33	8	81	80	4	31
34	4 25	38	83	3.5	38
36	5 25	39	84	0 4	2 5
37	2 75	11	85	1 8	11
39	5	20	86	1	9.5
40	7 5	33	87	3	35
41[ab]	3 17	15	88[a]	1 65	10
42	0.8	7	89	2	22

* Unit, one dollar. Prices are for December 17, 1937. Source: *Earnings Bulletin*, Standard Statistics Company, January, 1938, pp. 8, 10–11, 13, 16. Companies are numbered serially for the industrial classes covered: group 12-A, electrical equipment, numbers 1–14; group 19-A, machinery and machine equipment, numbers 15–51; group 19-B, agricultural machinery, numbers 52–60; group 27, office and business equipment, numbers 61–71, group 31, railroad equipment, numbers 72–90. Companies numbered as follows were excluded from the present tabulation, for reasons stated: 1, deficit; 5, earnings stated in per cent; 14, 31, 35, 38, 44, 51, 53, 57, 69, 82, no earnings stated; 46, 74, 81, 90, earnings include nonrecurrent element; 58, 68, earnings stated for less than year.

[a] For fiscal year ending other than December 31.

[b] Actual earnings. All other cases are estimates of the Standard Statistics Company.

[c] Both earnings and price divided by 2, for these companies, to bring figures within narrow range of variation.

venient to select the horizontal direction for that variable more likely to be regarded as "independent." The items in the correlation table are not values of the original variables (price and earnings per share, in the illustration), but frequencies (the number of companies having price and earnings as specified by the respective captions and stubs). For example, the original data are supposed given as in Table 88, a categorical table of the second

TABLE 89

CORRELATION TABLE FOR DATA OF TABLE 88

Y: Dec. 17, 1937 Price	X: Earnings per share, in dollars, lower limit of each interval																Totals of rows
	0	.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	
80-85																	1	1
75-80																		
70-75																		
65-70									1			1					1	3
60-65																		
55-60																		
50-55									1			1		1				3
45-50									1		1							2
40-45					1				1	1								3
35-40							1	1	1		1							4
30-35							1		2		1				1	1		6
25-30			1						3				1					5
20-25					3	2	1				1							7
15-20			1	1	2	1	3											8
10-15			1	3	1	5	1	1										12
5-10	1	6	3	3			1											14
0-5	3	1																4
Totals of columns	4	7	6	7	7	8	8	2	10	1	4	2	1	1	1	1	2	72

order, in which stubs indicate (by the assignment of serial numbers) classification according to companies, and captions distinguish the two members of each pair of associated items. Table 89 is derived from the categorical table by listing the frequencies of companies having earnings and price as indicated by the captions and stubs. In constructing this table the original data for earnings have been classified in 50-cent intervals, and those for

price in 5-dollar intervals, as indicated. Usually in a correlation table, as in this case, the class interval is not the same for both variables; and even the original units of measurement may differ, as would be the case if we were studying the correlation between price per share and number of shares sold on the exchange in a month.

Suppose that the variable represented by the captions is called X, and that the stubs indicate intervals of size for the variable Y. To each caption (interval of X) there corresponds a column of items which are the frequencies: each of these items is the number of those cases (companies), in the limited group for which the value of X is in the interval stated by the caption, and for which Y is in the interval indicated by the stub. Thus each column, taken with the stubs, constitutes a frequency series in which the variable is Y. Each such column is called a *Y-array* (also sometimes *vertical array*), and a particular column, for some caption X, is specified as the *Y-array of type X*. Likewise, each row is a frequency series in the variable X, and a particular row is called an *X-array of type Y*. The rectangle at the intersection of a column and a row is called a *cell*.

As each Y-array is a frequency series, it can be analyzed by the methods already studied for such series. It has a total frequency, a mean, a standard deviation, and a skewness. The values of these characteristics are likely to differ for different vertical arrays. In particular, the means of the several vertical arrays are likely to be different values of Y.

The scatter diagram.—Any pair of associated variates may be represented diagrammatically by plotting a point on a two-dimensional chart called a *scatter diagram* (Chart 90). This chart is plotted from the categorical data, given in Table 88. Similarly, the mean of each vertical array can be plotted at the appropriate point.[1] If all these means are equal, the mean points will lie along a horizontal line in the scatter diagram. Even if all the means of the Y-arrays are not equal, they may merely differ irregularly in such manner that the mean points appear to be scattered along a horizontal line. In either of these cases the diagram would indicate that there is no correlation, since in neither case do large

[1] These array means can, of course, be calculated by considering each vertical array as a separate frequency series, and applying the customary methods of computation elaborated in Chap. XI. With some ingenuity, however, the statistician can devise schemes for abridging the total work of calculating means for the whole batch of arrays.

values of X appear to have different mean values of Y than do small values of X.

If, on the other hand, these mean points appear to cluster along an inclined line, the diagram suggests that there is correlation. If such line is inclined upward to the right, correlation is *positive:* large values of X are accompanied by large mean values of

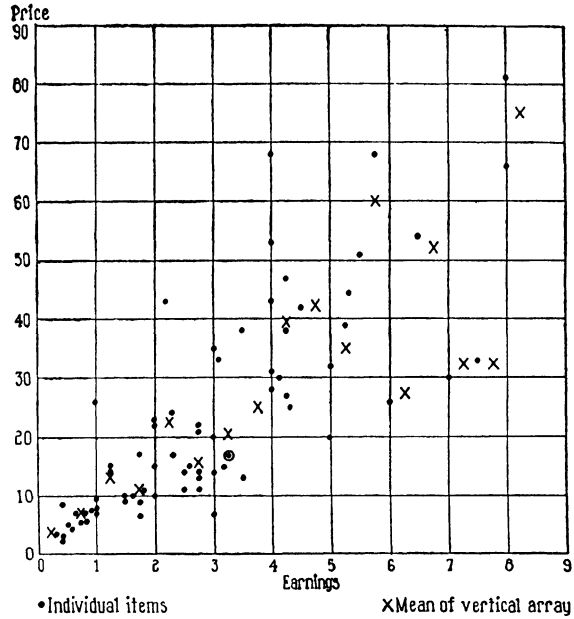


CHART 90.—Scatter diagram for the data on earnings per share and stock prices of selected companies in 1937.

(Data in Table 88, page 232.)

Y (this assumes X and Y increase toward the right and upward, respectively). If the line inclines downward to the right, correlation is said to be *negative.* The clustering may follow some curve, rather than a straight line; but, for the present, curvilinear correlation will be disregarded. When the clustering is along a straight line, regression is said to be *linear* (strictly, *rectilinear*).

THE LINES OF REGRESSION

The line which lies as near as possible (*nearness* to be defined as below, Chap. XVI) to the means of the Y-arrays is called the *line of regression of Y on X.* In a similar manner, the means of the rows —X-arrays—can be found, and the line which lies as near as

possible to these means is the *line of regression of X on Y*. Ordinarily the two lines are not identical, but intersect at a point called the *general mean* of the scatter diagram (indicated by *M* in Chart 91).

Frequently the means of the arrays are not actually computed, but an estimate of their position is made by inspection of the correlation table. This is often, as in the case of Table 89, but by no means always, sufficient to ascertain whether these means cluster

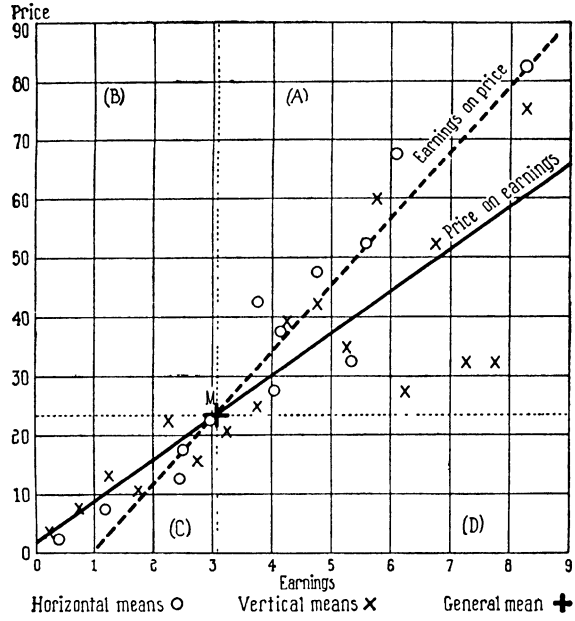


CHART 91.—Lines of regression for Chart 90.

along a line and whether such line is inclined. In this manner the correlation table in practice yields a preliminary estimate of the existence and extent of correlation. In making such visual appraisal of correlation from the correlation table, the observer should consider the actual size of the frequency within each cell and not merely the fact that a cell has some frequency. Such estimate is often facilitated by forming the correlation table with scores or crosses, one for each pair of variates entered in a cell, instead of a single number for each cell (Table 90). Such a table would often, in practice, be made as an intermediate step to forming Table 89 (see above, page 233). Observation of the correlation

table should lead to roughly correct preliminary conclusions on the questions whether any correlation exists, whether such correlation is positive or negative, high or slight, and whether the regression is of the straight line or other sort.

TABLE 90

SCORED COPY OF CORRELATION TABLE FOR DATA OF TABLE 88

Y: Dec. 17, 1937 Price $	X: Earnings per share, in dollars, lower limit of each interval																
	0	.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8
80-85																	1
75-80																	
70-75																	
65-70									1			1					1
60-65																	
55-60																	
50-55									1			1		1			
45-50									1		1						
40-45					1				1	1							
35-40							1	1	1		1						
30-35							1		//		1				1	1	
25-30			1						///				1				
20-25					///	//	1				1						
15-20			1	1	//	1	///										
10-15			1	///	1	/////	1	1									
5-10	1	//////	///	///			1										
0-5	///	1															

THE DISTRIBUTION OF FREQUENCY WITHIN EACH ARRAY

Actual computation of the means of the vertical and of the horizontal arrays generally affords more reliable information. Moreover, important supplementary knowledge is gained by study of the standard deviations and the skewness coefficients of the several arrays.[1] Distributions having the dispersions of all

[1] If σ is the same for all Y-arrays, the distribution is called *homoscedastic* in the variable X; and, if σ is the same for all X-arrays, the distribution is *homoscedastic* in the variable Y. If the skewness coefficient is zero for every array, the distribution is *homoclitic*.

vertical arrays equal to each other, the dispersions of all horizontal arrays equal to each other, the skewness zero for every array of both directions, and both regressions linear, are ideally favorable for correlation analysis. For such distributions the mathematical analysis of the following chapter yields most significant results; but, even if the distribution does not have all these properties but has linear regressions, such analysis is highly valuable. In practice, if regression appears to be rectilinear, the arithmetical analysis proceeds forthwith. It is occasionally important, however, to make this more thorough study of the anatomy of the correlation table itself in order to know in detail the dispersion and skewness of the several arrays.

CHAPTER XVI

THE COEFFICIENT OF CORRELATION

THE CORRELATION COEFFICIENT AS BASED UPON THE SCATTER DIAGRAM

We have seen in the preceding chapter that associated pairs of two variables, such as those given in Table 88, can be represented graphically by a scatter diagram (Chart 90). For every associated pair of variates there is one point on the scatter diagram, and the arrangement of these points indicates whether and

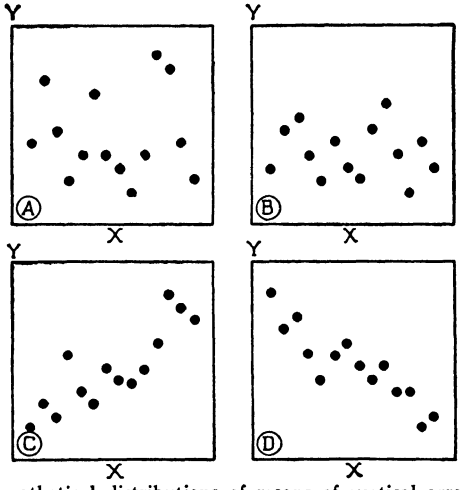


CHART 92.—Hypothetical distributions of means of vertical arrays, to indicate existence and nonexistence of correlation.

in what degree correlation between the two variables exists. If these points are scattered irregularly and with no systematic tendency to cluster along some line (or curve), as in Chart 92A, or if they cluster along a horizontal line, as in Chart 92B, there is no correlation. If the points cluster along a line inclined positively (upward to the right), as in Chart 92C, correlation exists and is positive; and, if the clustering is along a negatively inclined line, as in Chart 92D, correlation exists and is negative.

Suppose that the mean is calculated for each of the two variables: let M_x be the mean value of earnings and let M_y be the mean value of price for the data of Table 89. The corresponding point, M, on the scatter diagram (Chart 91) is called the *general mean.* If a horizontal and a vertical line—the *mean lines*—be drawn through M, the field of the scatter diagram is divided into four sections, A, B, C, and D. In section A both earnings and price are greater than their respective means, and in section C both are smaller than their respective means. In sections B and D, on the other hand, one variable is greater than its mean while the other is less than its mean. If the individual variates are expressed as deviations from their respective means—if, in other words, the positions of the points of the scatter diagram are specified by giving in each case the abscissa and ordinate with reference to the mean lines, these deviations are called x and y where

$$x = X - M_x$$
$$y = Y - M_y$$

Clearly x and y are both positive in section A, and both negative in section C, whereas one is positive and the other is negative in sections B and D.

If correlation exists and is positive, sections A and C will in general contain the greater portion of the plotted points; and, if it is negative, B and D will contain the greater portion of the points. Of course a rough measure of correlation might rest upon the percentage of points which fall in sections A and C, but such a measure would give as much weight to a point falling close to the mean lines as to a point falling anywhere in the section. It would not measure the tendency for the points to cluster about some line through the field, inclined say at about 45°. Such a measure would be subject to criticism in that many of the points could be shifted about, and yet leave the measure unchanged, whereas a good measure of correlation should take into consideration the exact position of every point, as well as the fact of its location in one of the four sections.

The products xy in sections A and C are all positive, and in B and D they are all negative. Evidently the location of a point in section A or C is indicated by a positive xy—called the cross product—for that point, and location in section B or D is indicated by a negative xy. If there is a greater tendency for points to cluster in A and C than in B and D, the sum of the positive cross products, therefore, will exceed the sum of the negative cross

products: the net total of all the cross products will be positive. If, on the other hand, there is a greater tendency to cluster in sections B and D, the net total of terms xy will be negative. When there is no greater tendency to cluster in A and C than in B and D (or *vice versa*)—when there is no correlation—the net total of terms xy tends to be zero. This net total of the terms xy, divided by the total frequency N, is called the *cross moment* (or *product moment*) and is a moderately good measure of correlation.

The cross moment is not an entirely satisfactory measure, however, for it is not appropriate as a means of comparing the degree of correlation for two or more problems each involving a pair of associated variables. The obstacle to comparison is the same as found in the case of measures of dispersion: the cross moment depends upon the dimensions of the two variables and the units in which those variables are expressed. This difficulty is removed through the conversion of x and y into standard units by dividing by the respective standard deviations. The cross moment as found for x and y expressed in standard units is called the *coefficient of correlation* and is designated by r.[1]

The *coefficient of correlation, r, is the arithmetic average of the products,* one product for each of the N associated pairs, *of the deviation of one variate,* from the mean of those variates for all N cases, *by the deviation of the associated variate,* from the mean of the associated variates for all N cases, deviations being measured in standard units.

CALCULATION OF r FROM THE CATEGORICAL TABLE

For example, columns 1 and 2 of Table 91 repeat the variates listed in Table 88; columns 3 and 4 give the deviations of these variates from their respective means, measured in the natural

[1] It is also called the *product moment coefficient of correlation,* because of its dependence upon the product moment. Moreover, because Professor Karl Pearson first fully developed it, it is frequently referred to as the *Pearsonian coefficient.*

The algebraic definition of the coefficient is $r = \Sigma uv/N$ where u and v are the variates expressed in standard units. Then, as

$$u = \frac{X - M_x}{\sigma_x}, \qquad v = \frac{Y - M_y}{\sigma_y}$$

$$r = \frac{\Sigma(X - M_x)(Y - M_y)}{N\sigma_x\sigma_y}$$

$$= \frac{\Sigma XY - M_y\Sigma X - M_x\Sigma Y + \Sigma M_xM_y}{N\sigma_x\sigma_y}$$

$$= \frac{\frac{1}{N}\Sigma XY - M_yM_x}{\sigma_x\sigma_y}$$

TABLE 91

CALCULATION OF CORRELATION COEFFICIENT BY DIRECT APPLICATION OF THE DEFINITION

Serial number	Earnings X	Price Y	$X - M_x$	$Y - M_y$	$\frac{X - M_x}{\sigma_x} = u$	$\frac{Y - M_y}{\sigma_y} = v$	uv		XY
							−	+	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
2	0 9	7.5	−2 06	−15 75	−1 0831	−0 8925		0 96667	6 75
3	0.65	7	−2 31	−16 25	−1 2145	−0 9208		1 11831	4 55
4	2.3	24	−0 66	0 75	−0 3470	0425	−0.01475		55 20
6	0 4	3	−2 56	−20 25	−1 3460	−1 1475		1 54453	1 20
7	0.3	3.5	−2 66	−19 75	−1.3985	−1 1191		1 56506	1 05
8	2.2	43	−0 76	19 75	−0 3996	1 1191	−0 44719		94 60
9	0.8	5.5	−2 16	−17 75	−1.1356	−1 0058		1 14219	4 40
10	2 5	14	−0 46	− 9 25	−0 2419	−0 5242		0 12680	35 00
11	2.75	21	−0 21	− 2.25	−0 1104	−0 1275		0 01408	57 75
12[a]	4	53	1 04	29 75	0 5468	1 6858		0 92180	212 00
13	2.	15	−0 96	− 8 25	−0 5047	−0 4675		0 23595	30 00
15	3.25	17	0 29	− 6 25	0 1525	−0 3542	−0.05402		55 25
16	1 25	15	−1 71	− 8 25	−0 8991	−0 4675		0 42033	18 75
17[a]	2 75	14	−0 21	− 9.25	−0 1104	−0 5242		0 05787	38 50
18	1 75	6 5	−1 21	−16 75	−0.6362	−0 9491		0.60382	11.38
19	4 25	27	1 29	3 75	0 6782	0 2125		0 14412	114 75
20	1	8	−1 96	−15 25	−1 0305	−0 8641		0 89046	8 00
21	3	20	0 04	− 3 25	0 0210	−0 1842	−0 00387		60 00
22	3 25	17	0 29	− 6 25	0 1525	−0 3542	−0.05402		55 25
23	5.5	51	2 54	27.75	1 3354	1 5725		2 09992	280 50
24	2.5	11	−0 46	−12 25	−0 2419	−0 6942		0 16793	27 50
25	3	7	0 04	−16 25	0 0210	−0 9208	−0.01934		21 00
26	1	7 5	−1 96	−15 75	−1 0305	−0 8925		0 91972	7 50
27	1.5	10	−1 46	−13 25	−0 7676	−0 7508		0 57631	15 00
28	1.5	9	−1 46	−14 25	−0 7676	−0.8075		0 61984	13.50
29	4.	28	1 04	4 75	0 5468	0 2692		0 14720	112 00
30[ab]	4 11	30	1 15	6 75	0 6046	0 3825		0 23126	123 30
32	3	14	0 04	− 9 25	0 0210	−0 5242	−0.01101		42 00
33	8	81	5 04	57 75	2 6498	3 2724		8 67121	648 00
34	4 25	38	1 29	14 75	0 6782	0.8358		0 56684	161 50
36	5 25	39	2 29	15 75	1 2040	0 8925		1 07457	204 75
37	2.75	11	−0 21	−12 25	−0 1104	−0 6942		0 07664	30 25
39	5	20	2 04	− 3 25	1 0726	−0 1842	−0.19757		100 00
40	7.5	33	4 54	9 75	2 3870	0 5525		1 31882	247 50
41[ab]	3.17	15	0 21	− 8 25	0 1104	−0 4675	−0.05161		47 55
42	0.8	7	−2 16	−16 25	−1 1356	−0 9208		1 04566	5 60
43	1.25	14	−1 71	− 9 25	−0 8991	−0 5242		0 47131	17 50
45	2	10	−0 96	−13 25	−0 5047	−0 7508		0 37893	20 00
47	5	32	2 04	8 75	1 0726	0 4958		0 53180	160 00
48[a]	4	68	1.04	44 75	0 5468	2 5358		1 38658	272 00
49[ab]	2.57	15	−0.39	− 8 25	−0 2050	−0 4675		0 09584	38 55
50[a]	6	26	3 04	2 75	1 5983	0 1558		0 24902	156 00
52	4 25	47	1 29	23 75	0 6782	1 3458		0 91272	199 75
54[ab]	0 53	5	−2 43	−18 25	−1 2776	−1 0341		1 32117	2 65
55[a]	4.3	25	1.34	1 75	0 7045	0 0992		0.06989	107 50
56[a]	8	66	5.04	42 75	2 6498	2 4225		6 41914	528 00
59[ab]	5.32	46	2 36	22.75	1 2408	1 2891		1 59952	244 72
60	7	30	4.04	6.75	2.1241	0 3825		0 81247	210 00
61	2 75	22	−0 21	− 1 25	−0.1104	−0.0708		0 00782	60 50
62	1.75	17	−1.21	− 6 25	−0 6362	−0 3542		0 22534	29 75
63	3.5	13	0 54	−10 25	0 2839	−0 5808	−0.16489		45 50
64[a]	5.75	68	2 79	44 75	1 4669	2 5358		3.71977	391 00
65	2.3	17	−0 66	− 6 25	−0 3470	−0 3542		0 12291	39 10
66	0.75	5.5	−2 21	−17 75	−1 1619	−1 0058		1 16864	4 12
67[a]	2.75	13	−0.21	−10.25	−0 1104	−0 5808		0 06412	35.75
70	0.6	4.5	−2.36	−18 75	−1.2408	−1.0625		1 31835	2.70
71	6.5	54	3.54	30 75	1 8612	1.7425		3.24314	351 00
72	4	43	1.04	19 75	0 5468	1.1191		0 61192	172 00
73[a]	1	26	−1.96	2.75	−1 0305	0.1558	−0.16055		26 00
75	3.1	33	0.14	9.75	0.0736	0.5525		0.04066	102 30

TABLE 91 (*Continued*)

Serial number	Earnings X	Price Y	$X - M_x$	$Y - M_y$	$\frac{X - M_x}{\sigma_x}$	$\frac{Y - M_y}{\sigma_y}$	uv		XY
							−	+	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
76	0 4	8 5	−2 56	−14 75	−1.3460	−0 8358		1 12499	3.40
77ab	1 75	9	−1 21	−14 25	−0 6362	−0 8075		0 51373	15.75
78	4 5	42	1 54	18 75	0 8097	1 0625		0 86031	189 00
79	2	23	−0 96	− 25	−0 5047	−0 0142		0 00717	46 00
80	4	31	1 05	7.75	0.5468	0.4392		0.24015	124.00
83	3 5	38	0 54	14.75	0.2839	0 8358		0 23728	133.00
84	0 4	2.5	−2.56	−20 75	−1 3460	−1 1758		1 58263	1 00
85	1 8	11	−1 16	−12 25	−0.6099	−0 6942		0 42339	19 80
86	1	9.5	−1 96	−13 75	−1 0305	−0 7792		0 80297	9 50
87	3	35	0 04	11 75	0 0210	0 6658		0 01398	105 00
88a	1 65	10	−1 31	−13 25	−0 6887	−0 7508		0 51708	16 50
89	2	22	−0 96	− 1 25	−0 5047	−0 0708		0 03573	44.00
Total	213 10	1,674 0					−1 17882	58 39838	6,875 17

$M_x = 2.96$, $M_y = 23.25$, $\sigma_x = 1.902$, $\sigma_y = 1.765$

$\frac{X - M_x}{\sigma_x} = u$, and $\frac{Y - M_y}{\sigma_y} = v$, then $r = \frac{uv}{N}$. Check $r = \left[\frac{\Sigma XY}{N} - M_x M_y\right] \div (\sigma_x \sigma_y)$

$r = \frac{58\ 39838 - 1.17882}{72} = 0.7947$, or $r = \left(\frac{6,875.17}{72} - 68.82\right) \div 33.5654 = 0.7945$

units; columns 5 and 6 give the deviations measured in standard units; and columns 7 and 8 give the negative and positive products obtained by multiplying together the associated pairs of variates in columns 5 and 6. The net total of columns 7 and 8, divided by the number of pairs of variates, is the coefficient of correlation.

In many practical problems the computation shown in this example, which applies the definition of correlation coefficient directly to the original data in the categorical form, could be followed satisfactorily. In such problems N should not be large, else the labor of calculation becomes excessive. On the other hand, r is less significant as a measure of correlation for distributions of small total frequency, and its more appropriate use is generally limited to instances with large total frequencies.

Even for the restricted use which can properly be made of r in connection with distributions of small N, the calculation should be attended by a careful examination of the distribution as a whole. One of the essential points is that regression should be approximately linear if r is to be significant for the distribution. The obvious, and generally most satisfactory, means of testing the distribution for linearity of regression is by constructing the correlation table or the scatter diagram, as described in the preceding chapter. Either of these operations yields at once a decisive answer to the question as to type of regression. A less effective test involves an inspection of the derived items of Table

91. If regression is approximately linear, two conditions should be fulfilled: the two items of each pair in columns 5 and 6 should have like signs, and should bear to each other approximately the same ratio for all pairs.

The direct computation scheme of Table 91 is unusual and must be replaced by a method utilizing data in the form of a correlation table. The two chief reasons for this are that data frequently are given in a correlation table and are not available in categorical form, and that the labor of the direct computation is so extensive that important economy of effort results from calculating r from the correlation table even in those cases for which categorical data are given. In respect to the second of these considerations, the actual construction of the correlation table should be regarded as an essential step in either method, as it is the best means of testing for linearity of regression.

DERIVATION OF r FROM A CORRELATION TABLE

When a computation is based upon the correlation table, an assumption analogous to that involved in the determination of the standard deviation of a grouped frequency series is made: every case which falls within a given cell is assumed to fall exactly at the center of that cell: the values of the two associated variates (measured from their means) for every actual case tabulated in a given cell are assumed identical with the x and y of the center of that cell. The errors involved in this assumption will, in general, cancel one another; and the net total error will not be large, especially if the cell dimensions are moderate, and if the distribution is one for which measurement of r is appropriate. Such net error may be disregarded in nearly all practical problems. The following discussion assumes also that the class intervals are of equal width for each variable, but the class interval for one variable is not necessarily equal to that for the other; and, indeed, the variables may be entirely different in nature (as height and weight), and hence certainly not measurable in the same units. Lastly, it is understood there are no "open ends"—no groups of the type "all under" or "all over." If the second and third assumptions are not fulfilled, r can be calculated, at least approximately; but such calculation involves slight modifications of the general method which need not be described.

From the data of Table 89 a modified scatter diagram (Chart 93) can be made, which differs from Chart 90 in that all the points belonging to a particular cell of Table 89 appear on the diagram at a single point, the abscissa and ordinate of which are the deviations

of the center of that cell from the general mean. The computation of r then consists, by analogy to the development of page 241, in finding the cross product for the center of each cell, summing all these individual cross products (each counted a number of times equal to the frequency in the given cell), dividing by N to yield the

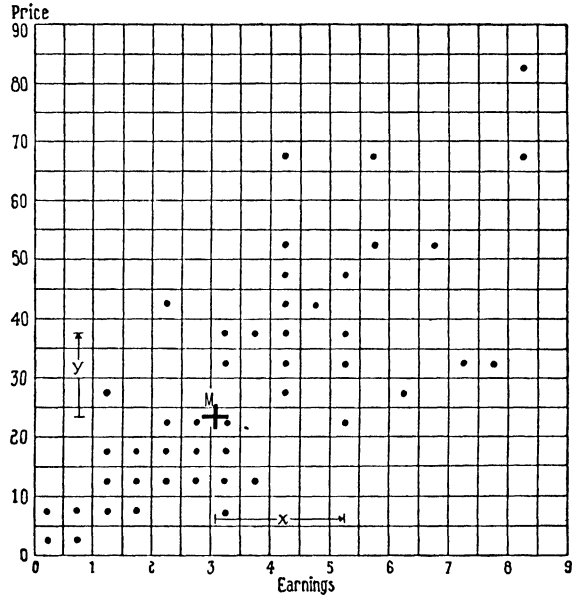


CHART 93.—Modified scatter diagram, in which paired varieties of Chart 90 are shown at centers of cells.

product moment, and dividing by the standard deviations for both x and y to yield r. These computations can be made from the correlation table, and appear in Table 92, wherein the lower figure in each cell is the weighted cross product.[1]

THE SKELETON METHOD

The labor of computation can be reduced very greatly by the use of a skeleton method. If x' and y' are the deviations of a given pair of variates from some arbitrary origin (usually chosen at the center of a cell near the general mean) and if d_x and d_y are the mean values of x and y as referred to that origin, r takes the form:

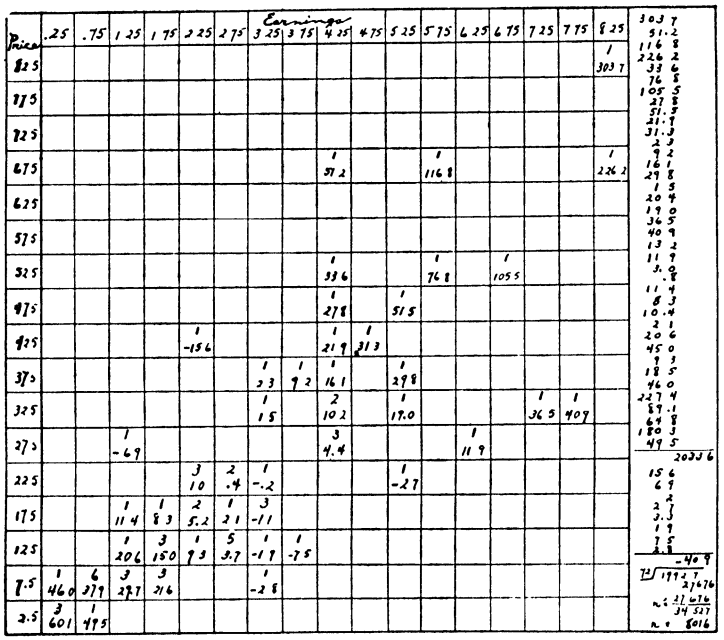
[1] Here the general mean is located, not from Table 91, but from the grouped frequency data: $M_x = 3.08$, $M_y = 23.75$.

$$r = \frac{\frac{1}{N}\Sigma x'y' - d_x d_y}{\tau_x \sigma_y}$$

This means that the product moment, $\frac{1}{N}\Sigma x'y'$, referred to the arbitrary origin should be computed, and then corrected by subtracting $d_x d_y$. The final step consists in dividing by the two standard deviations. Because of the fact that x' and y' are

TABLE 92

COMPUTATION OF CORRELATION COEFFICIENT FOR DATA OF TABLE 89 BY SPECIFIC-CELL METHOD



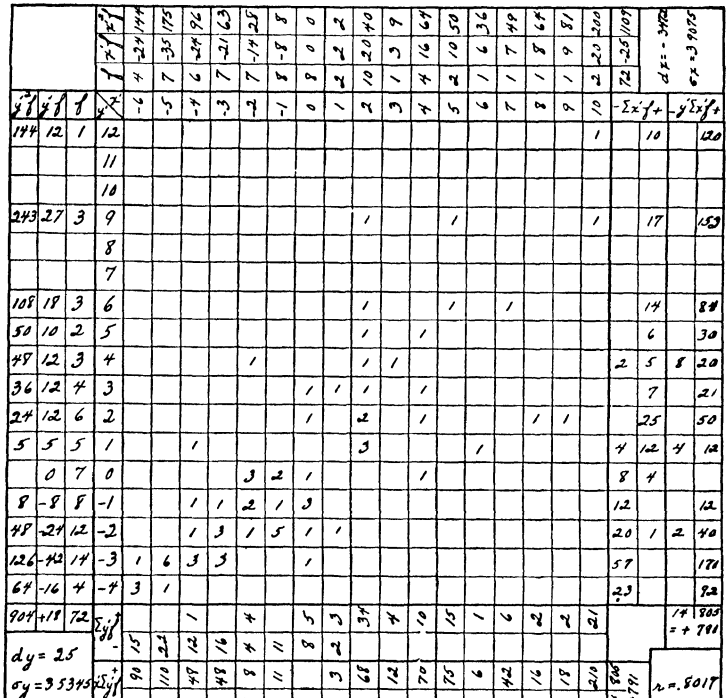
expressed in small integral numbers (it is understood that all measurements used here—in computing the product moment and the means and standard deviations—are in class intervals) the product moment referred to the arbitrary origin is much easier to compute than that referred to the general mean. Herein lies the great advantage in using the skeleton method.

The fact that x' has a single value for all cells of a given vertical array renders another simplification possible: the frequency of

each cell in that array can be multiplied by the appropriate y', these products can be added, and the result multiplied by x'. Similar operations for the other arrays lead to the items of the two bottom rows of Table 93, which, when totaled, yield 791. A corresponding group of computations can be made by use of the

TABLE 93

COMPUTATION OF CORRELATION COEFFICIENT OF TABLE 89 BY SKELETON METHOD



horizontal arrays. These steps appear in the columns at the right of Table 93, and yield a value for $\Sigma x'y'$ which is a check on that previously found.

THE LINE OF REGRESSION

The *line of regression of Y on X* was defined above (page 235) as the line which lies as near as possible to the means of the Y-arrays as plotted on the scatter diagram. The expression "as near as possible" is now defined as follows (see Chart 94). Sup-

pose the line already located, at AB. The deviation from that line of the mean, P_x, of a particular Y-array is then measured *vertically*, as M_xP_x. This deviation is squared, and multiplied by the total frequency f_x of the array; and the result is added to results similarly found for all the other Y-arrays. The aggregate thus derived is smaller than is the aggregate similarly derived for any line other than the line of regression of Y on X. In other words, the *line of regression of* Y *on* X is a line located in such manner that the sum of the squares of the vertical deviations of the means of Y-arrays from that line, each such square being

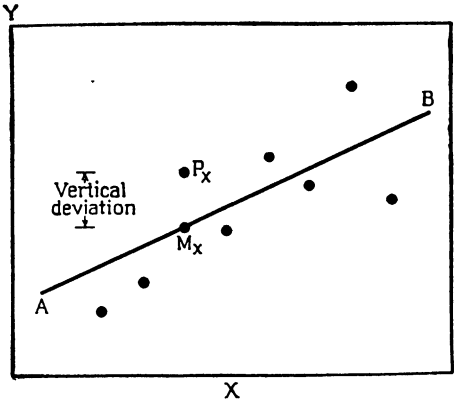


CHART 94.—Illustration of principle of location of regression line for a hypothetical distribution.

weighted with the frequency of the array, is minimum. The line of regression of X on Y is defined in analogous terms, except that here horizontal deviations of the means of the X-arrays are used. These definitions are specific applications of the notion of least squares.

It can be shown algebraically that the two lines of regression are not coincident (barring the exceptional case in which r is $+1$ or -1), that each line passes through the general mean, that the ordinate of any point on the line of regression of Y on X is r times the abscissa, and that likewise the abscissa of any point on the line of regression of X on Y is r times the ordinate—it being understood that *standard units* are used and the general mean chosen as the origin.[1]

[1] The algebraic forms of these statements where u and v are measured in *standard units* are

THE ERRORS OF ESTIMATE

Each line of regression provides an estimate of one variable from the other. For example, the line of regression of Y on X gives an estimate of Y for the given value of X. Except in the rarest cases, the actual value of Y for a given value of X will differ from the estimate afforded by the line of regression. Table 94 compares the actual variates with the estimated variates, and the errors are listed in columns 3 and 6. Each such difference is called an *error of estimate*, and is clearly identical with the "vertical deviation" mentioned above.[1] Hence a line of regression is a line located so that the sum of the squares of the errors of estimate is minimum.

For any one array (the following discussion will be confined to vertical arrays but, by obvious modifications, it is applicable to horizontal arrays) the ordinate of the line of regression of Y on X furnishes an estimate of the ordinate of the center of every cell in that array (Chart 95). The vertical deviation of the center of a

---

$$v = ru \qquad \text{and} \qquad u = rv$$

If the measurements are made in some other units, these equations become

$$y = r\frac{\sigma_y}{\sigma_x}x, \qquad x = r\frac{\sigma_x}{\sigma_y}y$$

and, if the measurements are made from an origin 0 instead of the general mean, the formulas are

$$Y - d_y = r\frac{\sigma_y}{\sigma_x}(X - d_x)$$

$$X - d_x = r\frac{\sigma_x}{\sigma_y}(Y - d_y)$$

where d_x and d_y are the deviations of the general mean from the given origin.

If this origin is the natural zero point and if the units of measurement are the natural units, these last formulas yield an estimate of Y (or X) in terms of X (or Y) as soon as the several constants M_x, M_y, σ_x, σ_y, and r are calculated from the data. Thus, for the example given in Table 88, the computations lead to $M_x = 3.08$, $M_y = 23.75$, $\sigma_x = 1.954$, $\sigma_y = 17.68$, $r = 0.8017$

$$Y = M_y + r\frac{\sigma_y}{\sigma_x}(X - M_x)$$

$$X = M_x + r\frac{\sigma_x}{\sigma_y}(Y - M_y)$$

and the two final equations reduce to $Y = 7.25X + 1.42$ and $X = 0.0887\,Y + 0.97$ in dollars as the equations of regression of Y on X and X on Y, respectively.

[1] Strictly, the vertical deviation discussed above is the deviation of the mean of the array from the line of regression. It can be shown, however, that, if the position of the line is defined in terms of deviations from it of the points representing actual pairs of variates, the same position of the line results. Hence, the idea of vertical deviation is extended to the positions of the individual points belonging to pairs of variates.

cell from the corresponding point on the line of regression is the error of estimate. If the regression is truly linear, the mean of every array falls exactly on the line of regression, and therefore the

TABLE 94

ACTUAL AND ESTIMATED PRICE OF STOCKS, FOR CASE OF TABLE 88*

Serial number	Actual Y	Estimated Y	Error of estimate	Serial number	Actual Y	Estimated Y	Error of estimate
2	7.5	7 9	0 4	43	14	10 5	− 3.5
3	7	6 1	− 0 9	45	10	15.9	5 9
4	24	18 1	− 5.9	47	32	37.7	5 7
6	3	4 3	1.3	48[a]	68	30.4	−37.6
7	3.5	3 6	0 1	49[ab]	15	20.0	5.0
8	43	17 4	−25 6	50[a]	26	44 9	18 9
9	5 5	7 2	1 7	52	47	32 2	−14 8
10	14	19 5	5 5	54[ab]	5	5 3	0 3
11	21	21 4	0 4	55[a]	25	32 6	7 6
12[c]	53	30 4	−22 6	56[a]	66	59 4	− 6 6
13	15	15 9	0 9	59[ab]	46	40 0	− 6 0
15	17	25 0	8 0	60	30	52 2	22 2
16	15	10 5	− 4 5	61	22	21 4	− 0 6
17[a]	14	21 4	7 4	62	17	14 1	− 2 9
18	6 5	14 1	7.6	63	13	26 8	13 8
19	27	32 2	5 2	64[c]	68	43 1	−24 9
20	8	8 7	0 7	65	17	18 1	1 1
21	20	23 2	3 2	66	5.5	6 9	1 4
22	17	25 0	8 0	67[a]	13	21 4	8 4
23	51	41 3	− 9 7	70	4 5	5.8	1 3
24	11	19 5	8 5	71	54	48 5	5 5
25	7	23 2	16 2	72	43	30 4	−12.6
26	7 5	8 7	1 2	73[a]	26	8 7	−17 3
27	10	12 3	2 3	75	33	23 9	− 9 1
28	9	12 3	3 3	76	8.5	4 3	− 4 2
29	28	30 4	2 4	77[ab]	9	14 1	5.1
30[ab]	30	31 2	1 2	78	42	34 0	− 8 0
32	14	23 2	9 2	79	23	15.9	− 7.1
33	81	59 4	−21 6	80	31	30 4	− 0 6
34	38	32 2	− 5 8	83	38	26 8	−11 2
36	39	39 5	0 5	84	2.5	4 3	1 8
37	11	21 4	10 4	85	11	2.7	− 8.3
39	20	37 7	17 7	86	9 5	8 7	− 0 8
40	33	55 8	22 8	87	35	23.2	−11 8
41[ab]	15	24 4	9 4	88[a]	10	13.4	3.4
42	7	7.2	0.2	89	22	15.9	− 6.1

* For units, source, and footnotes, see Table 88.

mean of the errors of estimate for any array is zero. If regression is not linear, the means of some or all arrays fail to lie upon the line, and the mean error of estimate is not zero for such arrays.

The deviations of the means of arrays from the line of regression may be erratic or systematic, as noted above (page 235). In the former case regression is approximately rectilinear but not well defined; in the latter case the regression is curvilinear. In either case the line of regression is not a good fit to the means of the arrays.

There are two aspects of *goodness of fit:* the extent of the tendency (1) of the means of the arrays to lie on or near the line of

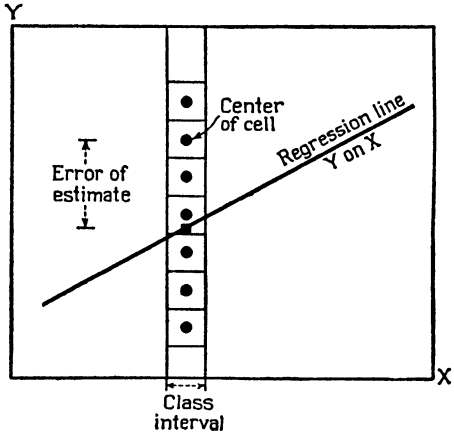


CHART 95.—Diagram of a particular error of estimate.

regression, and (2) of the variates in each array to cluster about the mean of that array. The first consideration concerns the appropriateness of the linear relation as representing correlation, and the second concerns the consistency of the distribution as a whole.[1]

Since the mean of the squares of the deviations from the regression line can be shown equal to

$$S_y^2 = \sigma_y^2 (1 - r^2)$$

r cannot be numerically greater than unity. The quantity S_y is the standard deviation of every Y-array only if the regression is

[1] A numerical test of the closeness, of the means of the several arrays, to the line of regression is given by

$$\zeta_{yx} = (\eta_{yx}^2 - r^2)$$

where

$$\eta_{yx} = \frac{\sigma_{my}}{\sigma_y}$$

and σ_{my} is the standard deviation of the means of arrays, each mean being weighted with the total frequency in that array. The nearer ζ is to zero, the more nearly rectilinear is the regression.

linear and the dispersions of all vertical arrays are equal. For cases which nearly meet these requirements, S_y is approximately equal to the standard deviation of any Y-array. The value of S_y, in such cases, affords a measure of the reliability of the estimate based upon the line of regression, for any array.

The correlation coefficient r and the regression coefficients $r\frac{\sigma_y}{\sigma_x}$ and $r\frac{\sigma_x}{\sigma_y}$ are characteristics of the correlation distribution: they are summary numbers which describe certain important features of the distribution. As in the case of other characteristics, the use of particular values of any or all of these numbers, found for a particular correlation table, as estimates of the values of the same characteristics for a larger group (population) of which only a portion is included in the sample covered by the data of the given correlation table, raises questions touched in the theory of sampling. The probable errors of these characteristics can be estimated, by methods analogous to that used in the case of the mean; and the interpretation to be placed upon each probable error in practice is as outlined in Chap. XIV.

CHAPTER XVII

FURTHER CORRELATION METHODS

THE CORRELATION RATIO

If the deviations of the means of the arrays from the line of regression are systematic rather than random, regression is curvilinear rather than rectilinear (as noted above, page 251). When regression is not truly rectilinear, r is not an entirely satisfactory measure of the degree of correlation. The means of the vertical arrays of Chart 91, while somewhat widely scattered, appear to have some systematic departure from the line of regression of Y on X. A slight tendency appears for the means to be above the line in the central part of the chart, and below it at the two ends. The coefficient r is not therefore an entirely satisfactory measure of correlation between the given stock prices and earnings. A broken line joining the means of the vertical arrays may be taken as the "curve of regression," and an appropriate measure of correlation is η_{yx} (eta for y on x).

$$\eta_{yx} = \frac{\sigma_{my}}{\sigma_y}$$

Here σ_{my} is the standard deviation of the means of arrays, each mean being weighted with the total frequency in the corresponding array.

The fraction η_{yx} (there is a symmetrically similar η_{xy}, defined in terms of the horizontal arrays) is called the *correlation ratio of y on x*. The sign of η is taken positive, and η_{yx} (or η_{xy}) is numerically greater than r, except when regression is exactly linear, in which case both η_{xy} and η_{yx} equal r. Except in that case, η_{yx} and η_{xy} need not be, and generally are not, equal.

The errors of estimate in any one array have a standard deviation, which is identical with the standard deviation of the array. The weighted average σ_{ay}^2 of the square of these standard deviations, the weights being the frequencies in the arrays, is related to the correlation ratio and the general standard deviation of the variates as follows:

$$\eta_{xy}^2 = 1 - \frac{\sigma_{ay}^2}{\sigma_y^2}$$

In like manner, if S_y^2 is the mean of the squares of the vertical deviations of the individual variates from the line of regression of y on x,

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$

The values of η_{yx} and η_{xy} for the data of Table 89 are 0.901 and 0.857.

SPURIOUS CORRELATION

The problem of the relation between earnings and stock prices might have been attacked from a point of view different from that of Chap. XVI. Instead of comparing the earnings per share with the price per share, one might compare total earnings available for common stock with total market value of common stock, for each company.

The value of r, calculated by methods of the previous chapter, for this new pair of variables would differ notably from that found for Table 89, but this new result measures the correlation between a quite different pair of variable phenomena. The r of Table 89 measures the correlation between prices and earnings *per share*, whereas the new value of r indicates relation between the two aggregates—total market value and total earnings. The degree of correlation for one case must not be inferred from the size of r for the other case.

Further light upon this important distinction is afforded by considering the notion of *spurious correlation*. A group of pairs of associated fractions may have a positive correlation although the numerators are quite uncorrelated. Moreover, if neither series of numerators is correlated with the series of divisors and if the numerators are not correlated with each other, the ratios will nevertheless be correlated unless all the divisors are equal. A correlation which results from the division of the two series of associated numbers by a third series is called *spurious correlation*.

For interpreting the results of a correlation study the above fact is significant. It is a consequence of the algebraic form of the correlation coefficient that the ratios may be correlated when the numerators are neither correlated with each other nor with the denominators; therefore, a correlation calculated for the ratios should not be taken as measuring the correlation between the numerators. Moreover, even if the numerators are correlated, the correlation between the ratios may be quite different. A correla-

tion coefficient is not *erroneous* because it is *spurious;* the error arises if the correlation between the numerators is inferred from that between the ratios. Thus, if one is interested in the correlation between earnings per share and price, there is no error in taking the r computed for the ratios—the r of Table 93. But it would be incorrect to state that value of r as measuring the correlation between aggregate earnings and market value.

PARTIAL CORRELATION

The doctrine of spurious correlation applies most clearly to cases in which the two series of numerators are uncorrelated and neither series of numerators is correlated with the series of divisors. In addition to the possible influence of a third variable upon the apparent correlation of two variables, arising from the formal algebraic relations, a real influence, due to the fact that each of the two given variables is correlated with a third, may exist. If both given variables are correlated with a common third variable, this accounts in part—in rare cases entirely—for the apparent correlation between the two given variables.

It is often desirable to find the degree of correlation of one variable with a second, independently of mutual correlation with a third. Such correlation between the first and second variables is called *partial correlation.* As the correlation of the first variable with the third can also be studied, independently of their mutual correlation with the second, there is likewise a partial correlation of the first variable with the third; and also there is a partial correlation between the second and third. Moreover, so far as the formal theory is concerned, the partial correlation of the first variable with each of any number of other variables can be found; but in practice there is seldom warrant for examining more than three or four mutually correlated variables. The complete determination of the relation of one variable to a group of several associated variables is a problem in *multiple correlation.*

The technique of determining partial correlation coefficients is very simple but involves extensive computations. Suppose the problem is to find the partial correlation of price per common share in December, 1937, Y, of the companies covered in Table 88, with earnings per share in 1937, X, and earnings per share in 1936, Z (data in Table 95). This examination is based upon the supposition that 1936 earnings constitute the "third variable"—that both the price and the earnings in 1937 are correlated with 1936 earnings. What is really in our minds is the question whether the

TABLE 95

EARNINGS PER SHARE IN 1937 AND 1936 AND PRICE PER SHARE DECEMBER 17, 1937 FOR COMPANIES COVERED BY TABLE 88*

Serial number	Price Dec. 17, 1937 *Y*	1937 earnings *X*	1936 earnings *Z*	Serial number	Price Dec. 17, 1937 *Y*	1937 earnings *X*	1936 earnings *Z*
2	7.5	0 90	1 30	43	14	1 25	1 32
3	7	0 65	0 20	45	10	2 00	2 28
4	24	2 30	1 99	47	32	5 00	3 35
6	3	0 40	0 49	48[a]	68	4 00	4 75
7	3 5	0.30	1 03	49[ab]	15	2 57	1.83
8	43	2 20	1 52	50[a]	26	6 00	4 58
9	5 5	0.80	1 03	52	47	4 25	2 27
10	14	2 50	2 13	54[ab]	5	0 53	0 14
11	21	2 75	2 86	55[a]	25	4 30	3 13
12[c]	53	4 00	2.84	56[a]	66	8 00	5 77
13	15	2 00	1 22	59[ab]	46	5 32	4 94
15	17	3 25	3 15	60	30	7 00	4 24
16	15	1 25	1 22	61	22	2 75	1 75
17[a]	14	2 75	2 74	62	17	1 75	1 39
18	6 5	1 75	0 99	63	13	3 50	1 62
19	27	4 25	3 27	64[c]	68	5 75	5 11
20	8	1.00	0.68	65	17	2 30	1 76
21	20	3.00	2 51	66	5 5	0 75	0 62
22	17	3 25	3 01	67[a]	13	2 75	1 77
23	51	5 50	5 23	70	4 5	0 60	0 59
24	11	2 50	1 81	71	54	6 50	5 06
25	7	3 00	1 74	72	43	4 00	2 95
26	7.5	1.00	1 02	73[a]	26	1 00	−1 36
27	10	1 50	1 36	75	33	3 10	2 64
28	9	1 50	1.08	76	8 5	0 40	0 00
29	28	4 00	3 81	77[ab]	9	1 75	−1 38
30[ab]	30	4.11	3 30	78	42	4 50	2 92
32	14	3 00	1 39	79	23	2 00	0 17
33	81	8 00	6 42	80	31	4.00	−1 38
34	38	4.25	2.97	83	38	3 50	2 95
36	39	5 25	4 27	84	2 5	0 40	−0.62
37	11	2 75	1 38	85	11	1 80	0 49
39	20	5 00	2 00	86	9.5	1 00	0 85
40	33	7.50	4 04	87	35	3 00	1 66
41[ab]	15	3.17	2 29	88[a]	10	1.65	0 80
42	7	0 80	0 73	89	22	2 00	1 76

* Units and source as in Table 88. For company 76, 1936 item is missing and has been taken as zero. For company 86, 1936 item is incomplete and is estimated herein for full year. See source for other qualifications of 1936 items.

1937 price may not be influenced by *both* 1936 and 1937 earnings. We are in fact regarding 1936 earnings as one of the possible "factors affecting" the 1937 price. The first operation consists in finding the three simple correlations: those between X and Y, X and Z, and Y and Z, which are denoted respectively by r_{xy}, r_{xz},

TABLE 96

CORRELATION TABLE FOR X AND Z OF TABLE 95

Z. Earnings per share, 1936	X: Earnings per share, 1937, in dollars, lower limit of each interval																Total of row
	0	.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	
6.5																		
6																	1	1
5.5																	1	1
5												2		1				3
4.5									1		1		1					3
4											1				1	1		3
3.5									1									1
3							2		3		1							6
2.5						2	2	1	3	1								7
2					1	1	1		1		1							5
1.5					4	4	2	1										11
1	1	2	3	3	1	1	1											12
.5		3	2	2														7
0	2	2		1	1													6
-.5																		
-1.	1																	1
-1.5			1	1					1									3
Total of column	4	7	6	7	7	8	8	2	10	1	4	2	1	1	1	1	2	72

and r_{yz}. These are found from Tables 89, 96, and 97. The partial correlation of X with Y, independently of Z, is then $r_{xy.z}$, given by

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$$

and the partial correlation of X with Z, independently of Y, is $r_{xz.y}$ given by

$$r_{xz.y} = \frac{r_{xz} - r_{xy}r_{yz}}{\sqrt{1 - r_{xy}^2}\sqrt{1 - r_{yz}^2}}$$

and, similarly, the partial correlation of Y with Z, independently of X, is $r_{yz.x}$ given by

$$r_{yz.x} = \frac{r_{yz} - r_{yx}r_{zx}}{\sqrt{1 - r_{yx}^2}\sqrt{1 - r_{zx}^2}}$$

The three simple correlations, and certain values derived in sub-

TABLE 97
CORRELATION TABLE FOR Y AND Z OF TABLE 95

Y: Dec. 13, 1937 Price	Z: Earnings per share, 1936, in dollars, lower limit of each interval																	Totals Y have
	-1.5	-1.	-.5	0	.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	
80-85																1		1
75-80																		
70-75																		
65-70													1	1	1			3
60-65																		
55-60																		
50-55									1					2				3
45-50								1					1					2
40-45							1		2									3
35-40							1		2			1						4
30-35	1								1	2		2						6
25-30	1									2	1		1					5
20-25				1			3	1	2									7
15-20						3	2	1		2								8
10-15				1	1	3	4	2	1									12
5-10	1			3	5	4	1											14
0-5		1		1	1	1												4
Totals Z have	3	1		6	7	11	12	5	9	6	1	3	3	3	1	1		72

sequent computations leading to the three partial correlations appear in Table 98.

The resulting values, 0.4780 and 0.2322, indicate the extent of the correlation of price with 1937 earnings (or 1936 earnings) apart from any influence of the mutual correlation of both with 1936 earnings (or 1937 earnings).

Just as a line of regression (with its equation involving x and y) was found as an approximate relation in the two-variable case

(page 249), so an equation of regression can be found among the three variables of the present problem. As in the two-variable problem there were two such equations—the regression of y on x and that of x on y—so here there are three *equations of partial regression.* One of these estimates x from y and z, one estimates y from x and z, and the other estimates z from x and y. The values of $r_{xy.z}$ and $r_{xz.y}$ are factors in the regression coefficients

TABLE 98

PARTIAL CORRELATION COEFFICIENTS OF TABLE 96 WITH THE COEFFICIENTS OF REGRESSION AND THE REGRESSION EQUATIONS

$M_x = 3\ 08$	$M_y = 23\ 75$	$M_z = 2\ 09$
$\sigma_x = 1\ 954$	$\sigma_y = 17\ 67$	$\sigma_z = 1\ 645$
$\sigma_{x.y} = 1\ 1082$	$\sigma_{y.x} = 10\ 5622$	$\sigma_{z.x} = 0\ 8890$
$\sigma_{x.z} = 1\ 0560$	$\sigma_{y.z} = 11\ 6959$	$\sigma_{z.y} = 1\ 0888$
$r_{y.z} = 0\ 7496$	$r_{x.z} = 0\ 8414$	$r_{x.y} = 0.8017$
$r_{yz.x} = 0\ 2322$	$r_{xz.y} = 0\ 6076$	$r_{xy.z} = 0.4780$
$= r_{zy.x}$	$= r_{xz.y}$	$= r_{yx.z}$
$b_{yz.x} = 0\ 2097$	$b_{xz.y} = 0\ 6184$	$b_{xy.z} = 0.4316$
$b_{zy.x} = 0\ 0195$	$b_{xz.y} = 0\ 5970$	$b_{yx.z} = 5\ 2942$
$X = 0.4316Y + 0.6184Z - 8.4630$	$Y = 5.2942X + 0.2097Z + 7.0056$	$Z = 0.5970X + 0.0195Y - 0.2119$

$b_{xy.z}$ and $b_{xz.y}$ which are parameters in the equation of *regression of* x *on* y *and* z

$$X - M_x = b_{xy.z}(Y - M_y) + b_{xz.y}(Z - M_z)$$

where

$$b_{xy.z} = r_{xy.z}\frac{\sigma_{x.z}}{\sigma_{y.z}}, \qquad b_{xz.y} = r_{xz.y}\frac{\sigma_{x.y}}{\sigma_{z.y}}$$

and

$$\sigma_{x.z} = \sigma_x\sqrt{1 - r_{xz}^2}, \qquad \sigma_{y.z} = \sigma_y\sqrt{1 - r_{yz}^2}, \qquad \sigma_{x.y} = \sigma_x\sqrt{1 - r_{xy}^2}$$
$$\sigma_{z.y} = \sigma_z\sqrt{1 - r_{yz}^2}$$

Just as the equation of the line of regression of X on Y (Chap. XVI) furnishes a means of estimating X from Y, the above equation of regression of X on both Y and Z affords an estimate of X when Y and Z are given. Similar regression equations can be developed for Y in terms of X and Z, and Z in terms of X and Y. Moreover, the partial correlation coefficients for cases with more than three variables can be expressed in forms similar to those given above for three variables.

The use of the partial correlation coefficients and the multiple regression equations rests upon the assumption that regression is approximately *planar*, the three-dimensional equivalent of *linear* (indeed, it is assumed that regression between each *pair* of variables is linear). It would be desirable to construct a three-dimensional graphic model, on the plan of the scatter diagram, to show whether the points representing triplets of associated variables cluster along a plane. This process would prove very laborious, and would not be available for problems of more than three variables. A makeshift test, which should *never* be omitted in practice, is the examination of the various two-dimensional scatter diagrams involved, to ascertain whether *all* of them show linear regression. For the three-variable problem, there are three such scatter diagrams: that for X and Y, that for X and Z, and that for Y and Z. If *any* of these shows curvilinear regression, the above type of partial correlation analysis is suspect.

The test of linearity of regression, developed in connection with the correlation ratio for the case of two variables, can be extended to three or more variables; but the algebraic forms involved and their satisfactory interpretation become rapidly more complicated and difficult as the number of variables is increased.

The considerable practical difficulty surrounding the test for linearity in multiple correlation is one of the reasons for the relative untrustworthiness of partial correlation coefficients. Furthermore, the attempt to appraise the partial dependence of one variable on each of several others is more hazardous than estimating the degree of relation between two variables. These are among the principal reasons for insisting that the methods of partial and multiple correlation be *not* applied unless the total frequency is large (the frequency in Table 95 is too small), and that the inferences drawn from the results of applying these methods be carefully guarded. The practical effect of these restrictions is that such methods have very limited application in economic statistics.

PART III

THE ANALYSIS OF TIME SERIES

CHAPTER XVIII

RELATIVES AND INDEX NUMBERS

COMMODITY PRICES

The methods developed above will be applied, in this and subsequent chapters, to the peculiar problems of time series. The first of these problems to be considered in this connection is that of the measurement of changes in commodity prices. Three of the many reasons for the prominence of commodity-price investigations in the study of economic time series are as follows: The measurement of changes in the cost of living is of wide public interest because it offers a means, however defective, of tracing fluctuations in average individual expenditure with a view to their comparison with variations in income. One ordinarily has in mind, in such measurement, the variations in the money cost of a fairly specific list of commodities and services in the roughly definite amounts which the average individual is supposed normally to require. Study of price maladjustment is a second major objective of price analysis, and is highly significant because the tendency for the prices of several important commodities or commodity groups to drift away from other commodity prices over a considerable time interval leads to serious alterations in the distribution of the income of the whole community and may operate to produce or prolong economic depression. A third reason for the large place of price studies in the problems of time series is that such studies furnish an approach to a determination of the purchasing power of money. The unit of value of commodities and services is itself a specific quantity of money. Therefore, changes over time in series expressed in monetary units are due in part to changes in the units themselves; and these latter changes must accordingly be estimated if other causes of price change are to be segregated. Because of the relative volume and reliability of data on wholesale commodity prices, change in the purchasing power of money has usually been inferred from the change in the general level of such prices.

For the present, changes in the purchasing power of the dollar due to purely monetary factors will be ignored (that is, the dollar

will be considered an unvariable standard of value), and variations in the quotations of a given commodity will be regarded as the true changes in value of that commodity.

PRICE RELATIVES

An effective device for comparing the price of a single commodity at one time with the price at another time is the *price relative.* A price relative is the ratio of the price at one time, called the *given price,* to the price at another time, called the *base price;* this ratio is usually multiplied by 100 to express it as a percentage. The base price may apply to an instant of time, as July 1, 1914, or may be an average for an interval of time, as the interval 1890–1899 or the year 1926. Such instant or interval of time is called the base; and although the base interval may be of any length as far as the formal definition of *relative* is concerned, there are important practical considerations tending to determine the interval. For the purpose of obtaining a representative average of price relatives it is desirable to select a base period that is relatively recent (see discussion on page 274). The period chosen as base should also be one in which the price relations are approximately normal. If the base period selected is to be a normal one in which no price maladjustments are present, it seems likely that a number of years should be included, for averages of prices over a period of years will minimize or eliminate entirely any maladjustments which happen to exist at any one period or over a short period of time.[1] It has seemed desirable here to follow the practice of the United States Bureau of Labor Statistics and use the average of prices in 1926 as the base. Column 1 of Table 99 gives a series of annual average prices, and if 1926 is taken as the base year the items of column 2 are the corresponding relatives. Thus each item of column 2 is the ratio of the corresponding item of column 1 to the 1926 price, and is expressed as a percentage of the average price in 1926.

If for each price relative the price in the preceding year (or interval) is used as base, the resulting relatives are those given in column 3 and are called *link* relatives. Thus, to get the link relative for a particular year, the price in that year is divided by the price in the preceding year. To distinguish them, the relatives shown in column 2 are called *fixed-base* relatives, as they are in terms of the unchanging base price of 1926.

[1] H. B. ARTHUR, "Wholesale Price Work of the Bureau of Labor Statistics," Washington Committee on Government Statistics and Information Services, p. 34.

By the use of the link relatives (column 3) it is possible to obtain a further set of relative prices, called *chain* relatives. For this purpose we take 100 as the chain relative for the year 1926; and for the following year the product of 100 by the link relative for the year 1927 (1.006), which gives 100.6 as the chain relative for 1927; and for 1928 the product of the chain relative for 1927 by the link relative for 1928 (1.136), which gives 114.3 as the chain relative for the year 1928, etc. In computing chain relatives prior to 1926, it is necessary to use the process of division

TABLE 99

YEARLY AVERAGE PRICE OF COTTON, MIDDLING, NEW YORK*

Year	Dollars per pound (1)	Relatives to 1926† (2)	Link relatives† (3)
1923	0 293	167 4	. . . .
1924	0 287	164 0	98 0
1925	0 235	134 3	81.9
1926	0 175	100 0	74 5
1927	0 176	100 6	100.6
1928	0 200	114.3	113 6
1929	0 191	109 1	95 5
1930	0 135	77 2	70.7
1931	0 085	48 6	63 0
1932	0 064	36.6	75.3
1933	0 087	49 7	135 9
1934	0 123	70 3	141 4
1935	0 119	68 0	96 8
1936	0 121	69 2	101.7
1937	0 114	65 1	94.2

* For source of data in column 1, see Table 100.
† Note: in this and the following tables calculations have been performed on a slide rule.

rather than that of multiplication. The chain relative for 1925 with 1926 as the base is 100 divided by the link relative 1926/1925 (100/.745) or 134.3; that for 1924 is the quotient of the chain relative for 1925 divided by the link 1925/1924 (134.3/.819) or 164.0, etc. The chain relative for a particular year is the product of the link relative for that year by the chain relative for the preceding year. Now, in consequence of the very definitions of fixed-base relative, link relative, and chain relative, it follows that the chain relatives are identical with the fixed-base relatives; the chain relatives are the items of column 2. This is true of the chain and fixed-base relatives of any single commodity, but

ordinarily no such simple relation holds between the chain and fixed-base index numbers pertaining to a group of commodities.

The series of fixed-base relatives affords a comparison of each price with the base price. Inspection of such series for several commodities, all with the same base interval, shows whether the

TABLE 100

PRICES OF COTTON, OATS, AND WHEAT AND THEIR PRICE RELATIVES ON 1926 AS A BASE*

Year	Prices			Relatives		
	Cotton, middling, New York (dollars per pound)	Oats, number 2 white, Chicago (dollars per bushel)	Wheat, number 2, hard, Kansas City (dollars per bushel)	Cotton	Oats	Wheat
	(1)	(2)	(3)	(4)	(5)	(6)
1923	0 293	0 439	1.112	167.4	102 1	74 3
1924	0 287	0 514	1 232	164 0	119 5	82 4
1925	0 235	0 467	1 670	134 3	108 6	111.6
1926	0 175	0.430	1 496	100 0	100 0	100.0
1927	0 176	0 497	1 372	100 6	115.6	91.7
1928	0 200	0 555	1 325	114 3	129 0	88.6
1929	0 191	0.486	1 180	109 1	113 0	78 9
1930	0 135	0 397	0 900	77.2	92 3	60 2
1931	0 085	0 278	0 606	48 6	64 7	40 5
1932	0 064	0 209	0 494	36 6	48 6	33 0
1933	0 087	0 294	0 724	49 7	68.4	48 4
1934	0 123	0 456	0 932	70 3	106 0	62 3
1935	0 119	0 417	1 040	68 0	97 0	69 5
1936	0 121	0 383	1 123	69 2	89 1	75 1
1937	0.114	0.437	1.201	65 1	101 6	80 3

* Source of data in columns 1–3; "Wholesale Prices, 1913 to 1928" (Bulletin 493), pp. 38–39, 52; "Wholesale Prices, 1929" (Bulletin 521), p. 19; "Wholesale Prices," 1930; (Bulletin 593), p. 42; and "Wholesale Prices," December issues, 1931 to 1937, Washington, U S. Bureau of Labor Statistics.

several price movements from the base interval are similar, and facilitates in some measure comparisons between minor movements. Table 100 presents, in the three left columns, the annual average actual prices for cotton, oats, and wheat, and, in the three right columns, the price relatives on the 1926 base. Some instructive comparisons are possible by direct examination of columns 4, 5, and 6; but, as for all tabulated material, no comprehensive view can be obtained from the data in this form.

GRAPHIC REPRESENTATION OF PRICE CHANGES

Chart 96 shows the data of columns 4, 5, and 6 on an arithmetic scale; and the main features of the movements, with their similarities and differences in both amount and direction, appear at a glance. It is evident that the plotting of the actual items of columns 1, 2, and 3 on an arithmetic scale would not be effective because of the wide differences in level of the curves. This

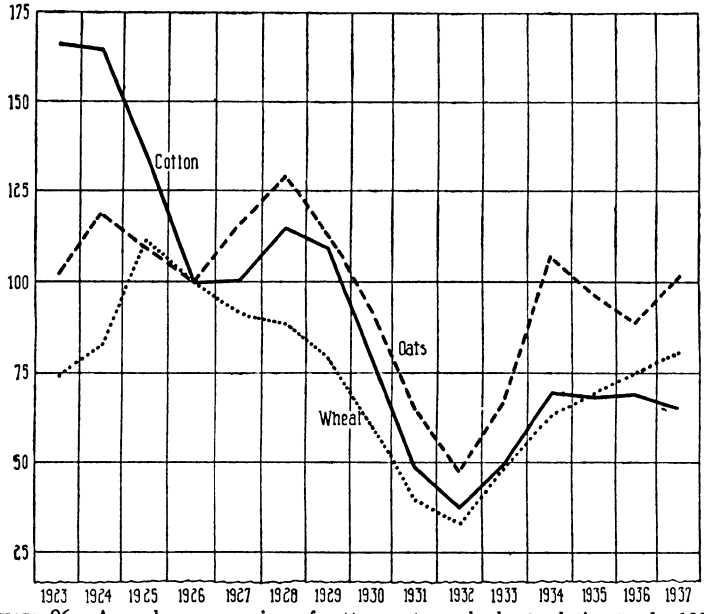


CHART 96.—Annual average prices of cotton, oats, and wheat relative to the 1926 price as 100 per cent.

(Data in Table 100.)

suggests one of the advantages of the price relatives: the level in the base interval is the same for all curves. Hence, appropriate selection of the base interval places any desired section of the several price curves on substantially a single level.

Chart 97 shows the items of columns 4, 5, and 6 on a ratio scale. This chart, in accordance with the discussion of Chap. VIII, emphasizes comparisons other than those made prominent by Chart 96; in Chart 97 the different rates of change, for the several commodities at any one time or for one commodity at various times, are most easily assessed.

The actual prices can be plotted on the ratio scale, and, as it is a property of such scale that vertical shifting of a curve does not

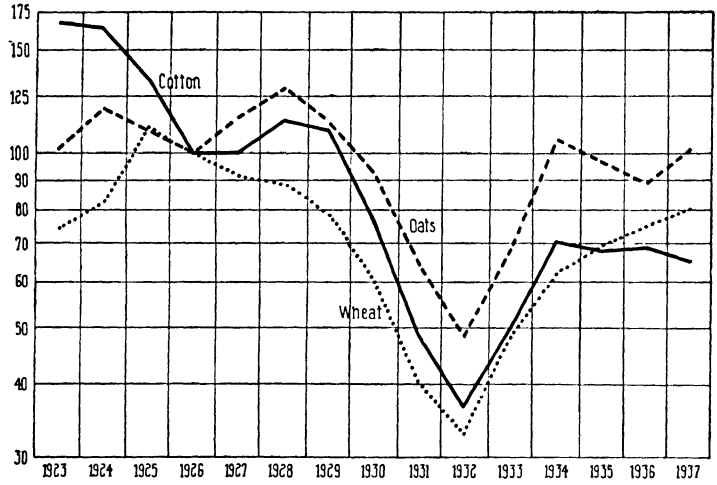


CHART 97.—Annual average prices of cotton, oats, and wheat relative to the 1926 price as 100 per cent.
(Ratio scale. Data in Table 100)

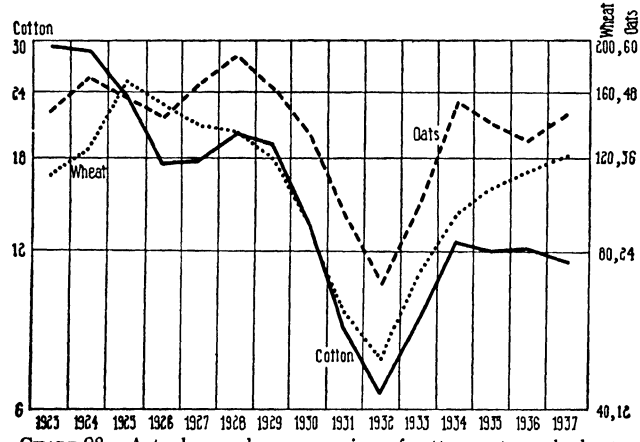


CHART 98.—Actual annual average prices of cotton, oats, and wheat.
(Units: Cents per pound for cotton, cents per bushel for oats and wheat. Ratio scale. Data in Table 100.)

change its form, the three curves can be placed approximately upon one level. This shifting is accomplished by using the scales indicated in Chart 98.

For the immediate purpose, Chart 98 suggests most significant conclusions. The striking fact is that the three curves do not usually move in the same direction—exceptional intervals are from 1925 to 1926 and from 1928 to 1934—and seldom, if ever, are the three rates of change identical. This is a characteristic feature of commodity prices, and shows that any one commodity could probably not serve as a sufficient representative of changes in the general price level. In order, therefore, to study changes in the price level as a whole, account must be taken of the variations in price of many commodities, and some summarization of all the individual movements must be used as an indicator of the general movement. Such a summarization is called an *index number*, or simply *index*, and it may be defined statistically as an average of the prices or of the price relatives of the individual commodities.

Index numbers will obviously vary depending upon the list of commodities for which averages are taken. It is not likely, for example, that an average of price movements over a period of time of the three commodities mentioned above—wheat, oats, and cotton—will move in exactly the same direction and by the same amount as an average of the price movements of pig iron, copper, and cement. The choice of the list of commodities to be included in the calculation of any index number, then, would seem to depend upon the *purpose* of the index number.

Some authorities (for example, Professor Irving Fisher) apparently believe that index numbers have one general purpose, namely, to measure changes in the purchasing power of money. In this case, an average must be obtained for all commodities and services which can be bought for money. Most authorities hold, however, that the purposes for which an index number may be constructed are varied. The investigator may wish to measure the movements in agricultural prices, in prices of textile goods, or in metal goods prices; or he may wish to ascertain the movements in relatively broad fields, such as wholesale prices, retail prices, prices of international commodities, or prices of those commodities sensitive to any impending change in the business cycle—the so-called sensitive prices. Upon the purpose for which the index number is constructed will depend the list of commodities included for which an average of price movements is to be obtained; incidentally, the purpose also determines to a considerable extent the type of average that is used.

Once the purpose is established, the list of commodities on which the index number is to be based must be selected. In the calculation of an index of wholesale prices, for example, all commodities and services which are purchased at wholesale theoretically must be included if the index number is to be significant. The task of obtaining all such wholesale prices is difficult, if not impossible. The present index number of wholesale prices published by the United States Bureau of Labor Statistics is based upon a list of 784 commodities, seemingly a very large list; and yet it is known that this group includes only a relatively small proportion of the total number of commodities sold at wholesale in the United States. Whether the index number of these 784 articles moves in exactly the same way as would the index number of all wholesale commodity prices is a question that cannot be answered with certainty. What is hoped is that the movement in the 784 prices is representative of the movement of all wholesale prices, and that an index number based on this list is representative of an index number that would be constructed from all prices if they were available.

Since it is generally not possible in the construction of an index number of wholesale prices to include all the commodities on which such an index should be based, a list of articles representative of the total must be selected. Such a list is sometimes, and somewhat loosely, called a *representative sample.* A representative sample is frequently made to include as many commodities as may be obtained, on the ground that the greatest possible completeness for the purpose in hand is desirable. More important than the above criterion, however, is the fact that the commodities chosen should be equally representative of all the types of movements which take place in wholesale commodity prices as a whole. It has been found, for example, that the movement in prices of raw materials is quite different from that which takes place in the prices of finished goods made from those same raw materials. An index number based on a list of commodities which consisted largely of raw materials, then, would move quite differently from an index number based largely on a list of prices of finished goods. Prices of goods which are used by the final consumer fluctuate in a different manner from those of intermediate goods which are sold to businessmen who subject them to further steps in the production process. A knowledge of these price interrelations is necessary if a representative list of commodi-

TABLE 101

RELATIVE 1927 PRICES OF SELECTED COMMODITIES ON THE 1913 AVERAGE AS BASE AND ON THE 1926 AVERAGE AS BASE*

Commodity*	Relative to 1913 average	Relative to 1926 average	Commodity*	Relative to 1913 average	Relative to 1926 average
Barley	132	119	Molasses	171	123
Corn	141	116	Oatmeal	142	114
Oats	132	116	Oleomargarine	137	98
Rye	164	110	Oleo oil	117	111
Wheat	156	92	Pepper	292	124
Cattle	149	133	Rice	116	80
Hogs	121	82	Salt	215	100
Sheep	178	101	Starch	127	100
Poultry	161	90	Sugar	135	106
Beans	153	112	Tallow, edible	111	93
Cotton	137	100	Tea	138	96
Eggs	133	90	Peas, canned	139	92
Apples	145	109	Tomatoes, canned	115	104
Lemons	136	140	Cottonseed oil	133	82
Oranges	161	119	Olive oil	126	111
Hay	121	83	Vinegar	173	104
Hops	128	91	Hides	106	139
Milk	181	103	Leather	110	112
Peanuts	153	110	Shoes, men	206	100
Tobacco	127	133	Gloves	251	100
Potatoes, white	217	71	Traveling bags	141	102
Wool	178	98	Blankets, cotton	176	86
Butter	149	107	Drillings, cotton	146	94
Cheese	170	111	Duck	124	98
Milk, evaporated	129	104	Flannel, cotton	199	101
Beef	152	118	Gingham	146	106
Lamb	176	100	Hosiery, cotton	198	98
Mutton	137	98	Muslin	173	93
Bacon	154	90	Percale	196	103
Hams	148	80	Print cloth	143	101
Pork	151	76	Sheeting	161	93
Poultry, dressed	160	86	Thread	187	100
Bread	167	100	Ticking	162	107
Cocoa beans	128	109	Underwear, cotton	174	91
Coffee	133	81	Yarn, cotton	142	98
Crackers, soda	215	100	Rayon	81	82
Fish, pickled cod	102	94	Silk	114	86
Flour	162	88	Silk hosiery	132	90
Apples, dried	161	98	Blankets, wool	172	97
Prunes	102	86	Flannel, wool	218	102
Bananas	143	90	Overcoating, wool	177	101
Glucose	153	94	Suiting, wool	213	97
Hominy grits	105	111	Shirts & drawers	176	93
Lard	117	86	Union suits	281	91
Corn meal, white	108	111	Dress goods, women	211	95

TABLE 101 (*Continued*)

Commodity*	Relative to 1913 average	Relative to 1926 average	Commodity*	Relative to 1913 average	Relative to 1926 average
Wool yarn	198	99	Stove, gas	154	101
Binder twine	137	94	Lumber, douglas fir	171	88
Burlap	122	106	Lumber, hemlock	160	99
Hemp	153	96	Lumber, oak	172	96
Sisal	177	84	Pine	163	93
Anthracite	227	98	Shingles	165	100
Bituminous coal	179	98	Brick	212	84
Coke	131	78	Cement	159	97
Petroleum, crude	138	68	Linseed oil	169	94
Gasoline	58	65	Shellac	237	139
Kerosene	123	73	Turpentine	145	67
Iron ore	125	100	Glass	155	91
Pig iron	123	96	Acetic acid	179	104
Bar iron	165	93	Nitric acid	133	101
Nails	145	96	Sulphuric acid	80	104
Pipe, cast iron	185	84	Ammonia	46	87
Saws, hand	189	100	Calcium chloride	163	100
Shovels	169	104	Indigo	78	100
Steel billets	127	95	Potash	153	96
Steel rails	143	100	Sulphur	82	99
Steel scrap	114	92	Alcohol	150	77
Steel sheets	141	96	Castor oil	139	104
Skelp	132	96	Epsom salts	212	94
Structural shapes, steel	125	94	Ether	188	94
Tin plate	154	100	Glycerine	126	90
Fence wire	169	96	Quinine	182	93
Wood screws	122	85	Nitrate of soda	101	98
Aluminum	108	94	Fertilizer, mixed	105	94
Antimony	156	78	Carpets	230	97
Copper	83	94	Cutlery	217	100
Lead	146	80	Glassware	139	91
Nickel	82	100	Plates	211	100
Quicksilver	276	127	Cattle feed	146	116
Silver	93	91	Newsprint	157	94
Tin	143	98	Woodpulp	126	93
Zinc	114	86	Rubber	46	78
Tractor	59	100	Lubricating oil	300	119
Automobile, Ford	73	110	Soap	135	93
Sewing machine	194	102	Starch, laundry	157	99
Stove, coal	188	98	Tobacco, smoking	148	100

* Compiled from "Wholesale Prices, 1913 to 1928," Washington, U. S. Bureau of Labor Statistics, Bulletin 493, Commodities appear in Table 9 of the Bulletin in order given above (certain commodities are omitted), and details may therefore be found in the source.

ties is to be chosen.[1] The foregoing brief discussion indicates that the selection of the list of commodities to be included for the construction of an index number for any specific purpose is based largely on individual judgment. Such individual choices, however, will not vary materially if the above criteria for selection are followed, and such variations in the list as do occur will not affect the index numbers which result to any considerable degree.

TABLE 102

FREQUENCY DISTRIBUTION OF PRICE RELATIVES FOR 1927 ON 1926*

Price relatives	Number of commodities	Price relatives	Number of commodities
Extremes[a]	8	96–97	13
118–119	4	94–95	14
116–117	3	92–93	12
114–115	1	90–91	12
112–113	2	88–89	2
110–111	8	86–87	7
108–109	2	84–85	4
106–107	5	82–83	4
104–105	7	80–81	4
102–103	5	78–79	3
100–101	27	76–77	2
98– 99	16	Extremes[a]	5

* Based on data in Table 101

[a] 140, 139, 139, 133, 133, 127, 124, 123; 73, 71, 68, 67, 65.

INDEX NUMBERS AS AVERAGES OF PRICE RELATIVES

Upon completion of the selection of commodities on which the index number is to be based, the next problem that arises is the choice of the appropriate average to measure price movements. The place of the theory of averages in the problem of price indexes is suggested by the following considerations: Table 101 shows the relative prices of a selected group of important commodities, with reference to two fixed bases, one recent and the other remote; and Tables 102 and 103 give the two corresponding frequency series. The frequency curves appear in Charts 99 and 100. Charts 101 and 102 show the deciles of each set of relatives as a system of diverging rays, and suggest the

[1] For more detailed discussion of these price interrelations and the problems that arise in index number work because of them, see W. C. MITCHELL, "Index Numbers of Wholesale Prices in the U. S. and Foreign Countries," Part I, Washington, U. S. Bureau of Labor Statistics (Bulletin 284), 1921, pp 39–48.

extent and nature of the dispersion in each series. Chart 103 compares the block diagrams for the two frequency series, and many of the leading features of the two distributions are evident at once. From each chart an impression can be formed of the extent of the dispersion and of the existence and nature of the skewness, and an estimate of the value of the mode can be made.

TABLE 103

FREQUENCY DISTRIBUTION OF PRICE RELATIVES FOR 1927 ON 1913*

Price relatives	Number of commodities	Price relatives	Number of commodities
Extremes[a]	24	144–145	3
188–189	3	142–143	6
186–187	1	140–141	3
184–185	1	138–139	5
182–183	1	136–137	5
180–181	1	134–135	2
178–179	4	132–133	8
176–177	5	130–131	1
174–175	1	128–129	3
172–173	4	126–127	5
170–171	3	124–125	4
168–169	3	122–123	4
166–167	1	120–121	2
164–165	3	118–119	0
162–163	4	116–117	3
160–161	6	114–115	4
158–159	1	112–113	0
156–157	4	110–111	2
154–155	4	108–109	2
152–153	6	106–107	1
150–151	2	104–105	2
148–149	4	102–103	2
146–147	4	Extremes[a]	13

* Based on data in Table 101

[a] 300, 292, 281, 276, 251, 237, 230, 227, 218, 217, 217, 215, 215, 213, 212, 212, 211, 211, 206, 199, 198, 198, 196, 194; 101, 93, 83, 82, 82, 81, 80, 78, 73, 59, 58, 46, 46.

If many such frequency series are formed and studied (care being taken that the base is not unduly remote), the distributions will be found in many instances nearly of the normal form and in almost no case more than moderately asymmetrical. Under such circumstances an average of all individual price movements obviously will be fairly representative of the movement in commodity prices as a whole. This conclusion, however, holds true only where the base period is not unduly remote; the dispersion of price relatives is not great when they are based on a relatively

recent base period, and the less the dispersion the more significant in general is the average. It is also obvious from the study of

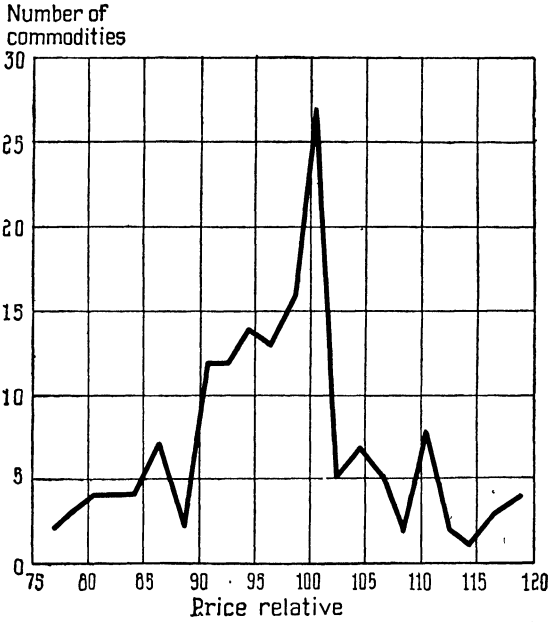


CHART 99.—Distribution of per cent ratios, of average price in 1927 to average price in 1926.

(Data in Table 102.)

Chart 100 that the more remote the base period the greater the dispersion and the less significant the average. In other words, the difficulty of obtaining an average representative of the total price movement increases with the increasing remoteness of the

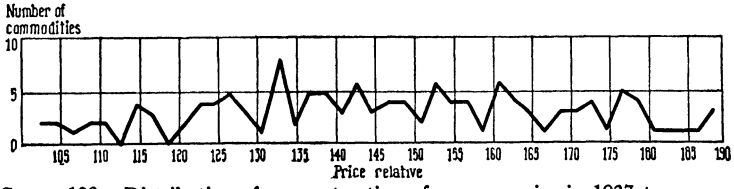


CHART 100.—Distribution of per cent ratios, of average price in 1927 to average price in 1913.

(Data in Table 103.)

base period on which the price relatives are computed. It may be concluded that in the construction of an index number over a

long period of time the base period should be changed occasionally so that the average of price relatives will be representative of the change in prices as a whole. Such change in the base period may be instituted every ten or fifteen years. Comparability between averages of price relatives constructed on different base periods may be obtained by calculating overlapping indexes at the time the change in base period is made.

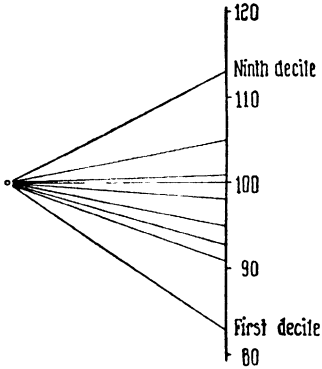


CHART 101.—Dispersion of price relatives for 1927 on the preceding year as base.
(Data in Table *V*, page 381.)

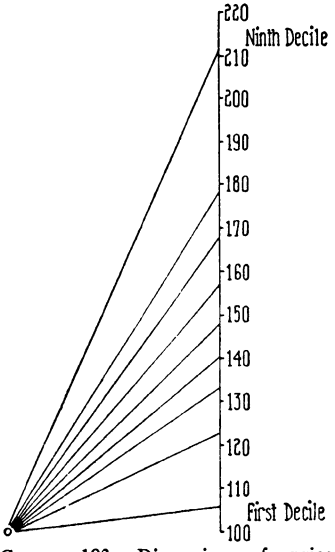


CHART 102.—Dispersion of price relatives for 1927 on a 1913 base.
(Data in Table *V*, page 381.)

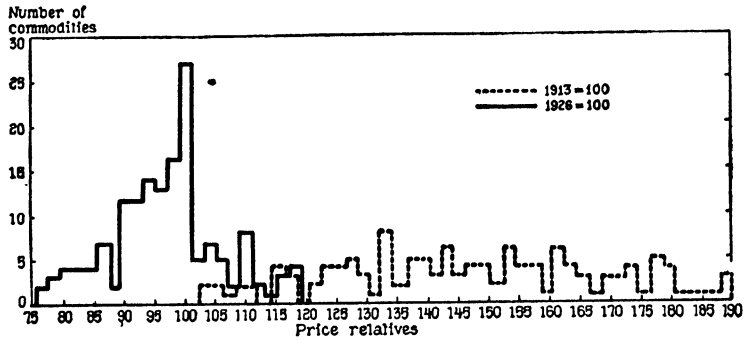


CHART 103.—Comparison of the distribution of price relatives for 1927 on the preceding year as a base and on 1913 as a base.
(Data in Tables 102 and 103.)

The fact that the frequency distribution of price relatives computed on a relatively recent base period approximates the

normal form tends to support one view of price movements: namely, that the movement of each commodity price is an approximation to the movement of the general price level but deviates from such movement by "errors" which are "accidental" in the sense of the theory of observations. In other words, the price movement of each commodity is assumed to yield a measurement of the movement of the general level of prices, but each such measurement may nevertheless be in error from causes which are as likely to produce positive as negative errors and more likely to produce small than large errors. Individual price movements are assumed to be independent of one another. The list of commodities from which the average movement is to be computed is considered to be a *random sample* of the total list of commodities theoretically available. The problem of the construction of an index number from this point of view becomes a problem in random sampling.

A priori considerations may suggest that the dominant force in the change in price of a particular commodity may well be the general change in the price level as a whole; experience shows that for a particular commodity the deviation frequently is not of the same sort as the "error" in the theory of observations. Thus it is often possible to establish the probability that some particular commodity will tend to increase in price more rapidly than commodities as a whole because of peculiar conditions affecting the market of the individual commodity. This situation may extend even to a whole group of commodities and lead to a cumulative departure of the prices of that group from the prices of commodities in general, a phenomenon called *price maladjustment.* In this view, each individual commodity evidently is not an equally good representative of the whole group, and the "error" of the individual commodity price is not strictly the sort of error implied in the theory of observations. The movements of individual prices are not independent of one another; and, as was pointed out previously, the index must be based on a list of commodities whose movements are representative of all the movements which take place. This second point of view of index numbers is more widely held than the first.

SIMPLE MEANS OF PRICES OR PRICE RELATIVES

With the selection of the commodities on which the index number is to be based and with the assurance that an average of the price movements of these individual commodities is fairly representative of the movement in the list as a whole, it becomes

necessary to choose that average which will result in the best index number. The use of different averages in the construction of simple index numbers may be illustrated by using the data of Table 104, which covers the price movements of five commodities of leading importance. The individual price movements can be averaged together in numerous ways to get the index number. A

TABLE 104

ANNUAL AVERAGE PRICES OF FIVE SPECIFIED COMMODITIES,* SUPPLEMENTED BY RELATIVE PRICES AND LOGARITHMS OF PRICES

	Wheat, number 2 hard, Kansas City	Hogs, fair to choice, heavy butchers, Chicago	Cotton, middling uplands, New York	Petroleum, crude, wells, Kansas-Oklahoma	Pig iron, basic, furnace
Actual price					
(1) 1926	1 496	12 336	0 175	1 884	18 548
(2) 1927	1 372	10 137	0 176	1 285	17 697
(3) 1928	1 325	9 628	0 200	1 203	16 664
(4) 1929	1 180	10 324	0.191	1.233	18 189
Price relatives, 1926 as base year					
(5) 1926	100 0	100 0	100 0	100 0	100 0
(6) 1927	91 7	82 2	100 6	68 2	95 4
(7) 1928	88 6	78 1	114 3	63 8	89 9
(8) 1929	78 9	83.7	109 1	65 4	98 1
Link relatives					
(9) 1927	91 7	82 2	100 6	68 2	95 4
(10) 1928	96 6	95 0	113 6	93 6	94 2
(11) 1929	89 1	107 2	95 5	102 5	109 2
Logarithms of prices					
(12) 1926	0 1749	1 0912	9 2430–10	0 2751	1 2683
(13) 1927	0 1374	1 0059	9 2455–10	0 1089	1 2479
(14) 1928	0 1222	9835	9 3010–10	0 0803	1 2218
(15) 1929	0 0719	1 0139	9 2810–10	0 0910	1 2598
Logarithms of price relatives, 1926 as base year					
(16) 1927	9 9624–10	9 9149–10	.0026	9 8338–10	9 9795–10
(17) 1928	9 9474–10	9 8927–10	0580	9 8048–10	9 9538–10
(18) 1929	9 8971–10	9 9227–10	.0378	9.8156–10	9 9917–10
Logarithms of link relatives					
(19) 1927	9 9624–10	9 9149–10	0 0026	9 8338–10	9 9795–10
(20) 1928	9 9850–10	9 9777–10	0 0554	9 9713–10	9 9741–10
(21) 1929	9 9499–10	.0302	9 9800–10	0.0107	0.0382

* Unit: dollars per bu. for wheat, dollars per 100 lbs. for hogs, dollars per lb. for cotton, dollars per bbl for petroleum, and dollars per gross ton for pig iron.

Source of data: "Wholesale Prices, 1913 to 1928" (Bulletin 493), pp 39, 46, 52, 137, 143; "Wholesale Prices, 1929" (Bulletin 521), pp. 19, 29, Washington, U. S. Bureau of Labor Statistics.

crude method is to represent general prices in each year by the sum—called the *aggregate*—of the five commodity prices in that year. Thus, the sum of the prices in row 1 (34.439) is the aggregate for 1926, and the corresponding figures for 1927, 1928, and 1929 are 30.667, 29.020, and 31.117. These figures might be regarded as index numbers for the respective years, but the chief objection to their use is obvious. Hogs and iron dominate the result because of the size of the units in which their prices are

quoted; but, had the price of wheat been stated in dollars per ton and that of cotton in dollars per bale, the results might have been widely different.

An aggregate is not properly an average at all. Division of the above results by 5 yields 6.888, 6.133, 5.804, and 6.223, and these are simple arithmetic averages of the actual prices in the respective years. The same objection evidently applies to these figures as to the aggregates themselves, and, hence, simple arithmetic averages of actual prices are unsatisfactory index numbers. Division of the aggregates for 1927, 1928, and 1929 by the aggregate for 1926 gives 89.0, 84.3, and 90.4 respectively; and these are the index numbers, according to this scheme of computation, for the three given years compared to the prices of 1926 as a base.

The chief objection to index numbers calculated by the aggregate method is that the results are determined largely by the units in which the prices of the several commodities are quoted. This influence can be eliminated entirely by using price relatives in the calculation of the index numbers, and the price relatives on 1926 as base for each of the commodities in the years 1926, 1927, 1928, and 1929 are given in rows 5, 6, 7, and 8, respectively of Table 104. Thus the average price of wheat in 1927 was 91.7 per cent of the average price in 1926; 91.7 is 100 times the ratio of $1.372, the price in 1927, to $1.496, the price in 1926; the other relatives are computed in like manner in accordance with the process defined above for cotton prices. All price relatives for 1926 are, of course, exactly 100. As soon as the price data have been put in the form of relatives, the nonhomogeneity due to units of price quotations is no longer present, and the five numbers for any particular year may be regarded as a series for which an average is to be obtained. The use of relatives in this way is, however, most satisfactory only if price relations are normal in the base interval.

Perhaps the most common method of summarizing the individual price relatives is to take their arithmetic average. Thus, the mean for 1927 of the relatives, given in row 6, is 87.6. Index numbers obtained in this way for the four years of the price record are: 100.0, 87.6, 86.9, and 87.0. An index number of this sort, which is widely used, is called a *simple arithmetic average of ratios*. Such an average, of course, is always affected by the presence of extreme items. Those commodities which fluctuate the most over a period of time determine largely the level of the mean.

OTHER SIMPLE AVERAGES

Instead of the mean of the five price relatives for any one year, the median might have been selected as the index number. On this plan, the index numbers for the four years would be: 100.0, 91.7, 88.6, and 83.7. These are seen to be strikingly different from the means for the same four groups of relatives. In general, when the number of items in a series is small—only five in this case—and those items are not grouped closely together, the mean and median differ quite widely from each other. The median, however, is not a good average to use when the number of items is as small as in this case since it is likely to be very erratic. If, on the other hand, the index number is based on a large number of price relatives, no fixed rule can be given to determine which average, the mean or median, should be used. In cases where a measure of the typical price movement is desired, a few extreme variations in widely nontypical commodities may render the mean less satisfactory than the median; but in cases, such as studies of changes in the purchasing power of money as a whole, where the irregular extreme price changes pertain to commodities which have an important bearing upon the problem under investigation, the elimination of the influence of these items by the median is not desirable (see Chap. XIX).

A third type of average which might serve as an index number is the mode. The mode cannot be used in a case in which there are only five commodities. No mode can be located for the series given in row 6. If, however, there were a large group of commodities, the price relatives for any one year might be grouped according to their sizes—say in class intervals 1 per cent in width—and the resulting frequency series would probably have a mode which might be taken as the index number for the year. Chapter XI showed that the statistical determination of the mode is at best an approximation; this fact, supplemented by certain theoretical objections, explains the general neglect of the mode as an index number.

An additional type of average which plays a considerable role in the construction and use of index numbers is the *harmonic mean*, which is defined as the reciprocal of the arithmetic mean of the reciprocals of the individual items. For example, the harmonic mean of the five items in row 6 is found by dividing each of the five items into 1, then getting the arithmetic average of these

reciprocals, which is 1.1631, and dividing this into 100, giving 86.0 per cent. The four index numbers calculated on this plan are: 100.0 86.0, 83.8, and 84.4. The significance of the harmonic mean is rather obscure, but a more concrete meaning appears by observing that the reciprocal of the number of dollars per pound is the number of pounds per dollar. Hence, in effect, the harmonic mean of the prices rests upon the arithmetic mean of the quantities obtainable per dollar at those prices.

The geometric mean of the five items of row 6 is the 5th root of the product obtained by multiplying all five price relatives

TABLE 105

INDEX NUMBERS OF PRICES FOR THE FIVE COMMODITIES, LISTED IN TABLE 104, WEIGHTED EQUALLY*

Type	Method	1926	1927	1928	1929
Arithmetic	Fixed-base	100 0	87 6	86 9	87 0
	Chain	100 0	87 6	86 4	87 0
Geometric	Fixed-base	100 0	86 8	85.4	85.7
	Chain	100 0	86 8	85.4	85 7
Harmonic	Fixed-base	100 0	86 0	83 8	84 4
	Chain	100.0	86 0	84 4	84 5
Median	Fixed-base	100 0	91 7	88 6	83.7
	Chain	100 0	91 7	87 1	89 3
Aggregate	Fixed-base	100 0	89 0	84 3	90 4
	Chain	100 0	89 0	84 3	90 4
Mode[a]	Fixed-base	100 0	. .	. .	
	Chain	100 0		.	

* Unit: per cent of 1926 average
[a] For so few as five items the mode is indeterminate.

together. By an elementary principle of logarithms, the logarithm of the geometric mean is equal to one-fifth the sum of the logarithms of the five relatives. The logarithms of these relatives for 1927 are given in row 16, and the geometric average computed by the use of these logarithms is 86.8. Similarly, geometric index numbers can be obtained for the other years, and the four results are: 100.0, 86.8, 85.4, and 85.7. These numbers are manifestly quite different from those obtained by the use of the mean, the median, and the harmonic average. In fact, the geometric index lies between the arithmetic and harmonic indexes.

FIXED-BASE AND CHAIN INDEXES

All the index numbers thus far discussed have been fixed-base index numbers. Each type of average used in getting fixed-base numbers may be used to get the corresponding series of chain index numbers. The process consists in computing link indexes for each year and then multiplying them together successively to get the chain indexes after the manner described above for getting chain relatives. For this purpose link relatives and their logarithms are shown in rows 9 to 11, and 19 to 21. The link relative for 1927, the first year after the base year, is always the same as the corresponding fixed-base relative. Using the mean as the type of average, the link index for 1927 is 87.6, for 1928 it is 98.6, and for 1929 it is 100.7. The corresponding chain indexes for the four years are: 100.0, 87.6, 86.4, and 87.0. A similar procedure may be followed, using each of the other types of averages, to get chain indexes. Table 105 shows that, excepting the geometric averages and the aggregates, the chain indexes are ordinarily quite different from the corresponding fixed-base indexes computed with the same type of average. Further observations on the properties of chain indexes will be presented in the critical survey of index numbers in the following chapter.

CHAPTER XIX

WEIGHTED INDEX NUMBERS

THE NECESSITY OF WEIGHTING

The index numbers discussed in the preceding chapter vary depending upon the type of average used. One important feature of these indexes is that they are obtained by taking simple averages of the prices or price relatives. Theoretically each commodity is just as important as any other in the computation of the index number. Actually, the type of average employed resulted in the assignment of a weight to each commodity which was not based upon the relative importance of that commodity, a phenomenon referred to as *implicit* weighting in contrast to *explicit* weighting, in which the commodities are weighted according to their relative importance for the purpose in hand. Strictly, this statement applies only to those indexes computed from price relatives; in the case of an index number resting upon the aggregate of actual prices, each higher priced commodity has greater influence upon the result than a lower priced commodity. The five commodities represented in Table 104 are of great importance in the industrial life of the nation, and, of course, a simple average of their price relatives might reflect with fair accuracy movements in the general level of prices of raw materials. Ordinarily the commodities used in constructing an index do not constitute so homogeneous a group and are by no means of so nearly equal importance as these five basic commodities. On the contrary, the group of commodities used in the construction of an index in general contains some, such as coal, cotton, and corn, the importance of which is many times greater than that of others, such as rye flour, tomatoes, and tin. The obvious objection to a simple average of the price relatives is that theoretically the less important commodities of the group have as much influence on the result as the more important commodities.

The situation is scarcely better with the aggregate of actual prices; for, although the influence on the result is not the same for all commodities, the differences in weight depend merely on the prices themselves and are likely to have no logical relation to the

true importance of the several commodities. Thus iron, which is of very great importance, would have less bearing on the index number than would diamonds, which have little importance, because the price of iron per ton is lower than that of diamonds per carat.

A method, somewhat arbitrary, of meeting the difficulty is to duplicate the price quotations of the more important commodities. Thus, less important constituents of the group, such as rice and lead, may be represented by a single quotation each; and more important goods, such as steel and wheat, may be represented by two or three or even more quotations each. The two quotations for steel probably comprise on this plan the prices of two slightly different grades, or of the same grade in different markets, but the effect is to enter steel twice in the computation of the index. This method of arbitrary weighting is applicable to indexes which are averages of relatives and to those which are aggregates of actual prices. If it is used with care and in the light of established facts concerning the comparative importance of the several commodities, the results obtained are likely to be very satisfactory. Several dependable published index numbers have been constructed on this plan.

THE PURPOSE OF AN INDEX AS A DETERMINANT OF THE WEIGHTING SYSTEM

A more precise way of taking account of the comparative importance of the several commodities is desirable, and the device used is the weighted average. This device has greater flexibility than the method of arbitrary weighting, for the weights used in constructing the average can be expressed in decimals and devised to represent quite accurately the relative importance of the constituent commodities. In actual practice there is a wide range of choice in the selection of weights. This is due partly to the fact that the relative importance varies according to the purpose for which the index is designed, and partly to the inherent difficulty of measuring the relative importance of the several commodities.

That the purpose for which an index is designed regulates its construction and, hence, its numerical value, is a point frequently ignored. Such oversight is particularly frequent in the use of published indexes, which are often cited for purposes entirely different from those in view when they were made. For example, an index of agricultural prices constructed from the point of view

of the consumer may be quoted as indicative of agricultural prosperity. The existence of several so-called general-purpose index numbers adds to the misunderstanding in this direction, for it is not made clear that a general-purpose index number is by no means an index number for manifold special purposes.

Most of these general-purpose indexes are, in fact, designed to measure the general level of prices, and are only properly available for special purposes which can be served adequately by measuring changes in the general price level. Although the following discussion primarily concerns such indexes of the general level of prices, the methods discussed are in large measure adaptable to the construction of more specialized indexes. Emphasis should be given, nevertheless, to the point that the special problems are distinct, and that appropriate special indexes are seldom identical with the general price index.

The measurement of the relative importance of the commodities for a given purpose, such as the construction of an index of the general price level, is difficult. Under one point of view of index numbers developed in the preceding chapter, that a change in price of a particular commodity is considered as an "observation" of the change in the general level of prices, the assignment of different weights to different commodities implies that some give better "observations" than others. Acceptance of this view may lead to several systems of weights, each determinable with more or less precision by statistical processes.

VALUES IN EXCHANGE AS WEIGHTS

The study of the weighting of indexes of general prices, however, has led to an entirely different view of the nature of index numbers. Instead of each commodity price being regarded as an observation on general prices, each commodity is considered as an element in the aggregate of commodities and services exchanged for money. The importance of a commodity in such an aggregate is measured by its total money value in exchange. This leads to the principle of weighting according to value in exchange. Even if the quantity entering in trade could be measured perfectly, such quantity is quite different from the quantity exchanged because of the different turnover rates for different commodities. Because of the paucity of data on the turnover of commodities, this consideration is ignored: quantity entered in trade is assumed in place of quantity exchanged. In most cases, then, the value in exchange is taken to mean the value of the total quantity produced

(sometimes with a correction for net excess of imports over exports), at the average price for the year. In Table 106 the annual production data are presented for each of the five commodities for the period 1926–1929. The value in millions of dollars for wheat in 1926 (see Table 106) is the product of 833.5 million bushels by the price, $1.496 per bushel, namely 1246.9

TABLE 106

ANNUAL AVERAGE PRICES OF THE FIVE COMMODITIES OF TABLE 104 SUPPLEMENTED BY FIXED-BASE PRICE RELATIVES AND PRODUCTION DATA*

	Wheat	Hogs	Cotton	Petroleum	Pig iron
Actual Price[a]					
(1) 1926	1.496	12 336	.175	1 884	18.548
(2) 1927	1 372	10 137	176	1 285	17 697
(3) 1928	1 325	9 628	.200	1 203	16 664
(4) 1929	1.180	10.324	.191	1 233	18 189
Production					
(5) 1926	833 5	157 9	17 98	770 9	39 07
(6) 1927	874 7	166 2	12 96	901 1	36.23
(7) 1928	913 0	183 1	14 48	901 5	37 84
(8) 1929	800 6	180 6	14 57	1007 3	42 49
Price relatives, 1926 as base year					
(9) 1926	100 0	100 0	100 0	100 0	100 0
(10) 1927	91 7	82 2	100 6	68 2	95 4
(11) 1928	88 6	78 1	114 3	63 8	89 9
(12) 1929	78 9	83 7	109 1	65 4	98 1

* Unit: all figures in millions: bu. for wheat, cwt. for hogs, bale of 500 lbs for cotton, bbl. of 42 gallons for petroleum, and gross ton for pig iron.

Production data from the following sources:

Wheat—"Agricultural Statistics, 1936," Washington, U. S. Department of Agriculture, 1936, p. 6.

Hogs—Production equals total number of hogs slaughtered multiplied by the average live weight at Chicago. Total slaughterings appear in "Agricultural Statistics, 1936," p. 224, average live weight at Chicago computed from the monthly averages which appear in "Yearbook of Agriculture, 1935," Washington, U. S. Department of Agriculture, 1935, p. 604, for monthly averages January to September, 1926, and in "Agricultural Statistics, 1936," p. 223, for remainder of period.

Cotton—"Agricultural Statistics, 1936," p. 76.

Petroleum—"Minerals Yearbook, 1937," Washington, U. S. Department of Interior, 1937, p. 1009.

Pig iron—*Iron Age*, January 3, 1935, p. 276.

[a] See footnote to Table 104.

million dollars; and that for hogs is the product of 157.9 by 12.336, namely 1947.9 million dollars.

The question immediately arises as to the time for which such value weights should be determined. An index number compares one year with another, and a series of index numbers compares several years with a single base year or with each other. For which of the years in the series should the values be determined in order to obtain the weights? One method consists in selecting

the values pertaining to a particular year and using them (or numbers proportional to them) as weights in calculating the indexes for all the years; this process is called *constant* or *fixed* weighting. An alternative method makes use of a special set of weights depending on the values in the "given" year; this process is called *variable* weighting.

Whenever fixed weights are used, some one year, or period, must be selected as that for which values in exchange are to be determined. Two obvious possibilities are the base year and the most recent year. If the problem pertains to indexes which compare more than one year with a fixed-base year, the use of given year weights does not yield fixed weights because the values in exchange vary from year to year. When constant weights are used, they are usually proportional to the values in the base year. Values in some other fixed year, or average values over a period of years, however, can be used. One might use quantities as given in a certain census year with the prices in some other year, as is done in the present wholesale price index of the United States Bureau of Labor Statistics.

THE CALCULATION OF WEIGHTED INDEXES

Let it be assumed that indexes are to be derived from the data of Table 106, with base-year values as weights. Then, in the

TABLE 107

CALCULATION OF THE WEIGHTED FIXED-BASE HARMONIC INDEX OF PRICES FOR FIVE SPECIFIED COMMODITIES FOR THE YEARS 1927, 1928, AND 1929*

Commodities	1926			1927		1928		1929	
	Prices p_0	Productions q_0	Weights p_0q_0	Prices p_1	$p_0q_0\frac{p_0}{p_1}$	Prices p_2	$p_0q_0\frac{p_0}{p_2}$	Prices p_3	$p_0q_0\frac{p_0}{p_3}$
Wheat	1 496	833.5	1246 9	1 372	1359 7	1 325	1407 4	1 180	1580 3
Hogs	12 336	157 9	1947 9	10 137	2369 6	9 628	2494 1	10 324	2327 2
Cotton[a]	0 175	8990 0	1573 2	0 176	1563 8	0 200	6376 4	0 191	1442 0
Petroleum	1 884	770 9	1452 4	1 285	2129 7	1 203	2276 5	1 233	2220 9
Pig iron	18 548	39 07	724 7	17 697	759 6	16 664	806 1	18 189	738 8
Total	. .	. .	6945 1	.	8182 4		8360 5	...	8309 2
Weighted index	..	.... .	100 0	...	84 9	.. ..	83 1	.	83.6

* The weights are proportional to the values in exchange in 1926. Index for 1926 = 100.
[a] Price and production in pounds.

manner indicated above, values are found for all of the commodities in the base year; and these value figures, or any set of numbers proportional to them, are the constant weights. To get the corresponding index number for 1927, assuming that the mean

TABLE 108

CALCULATION OF THE WEIGHTED FIXED-BASE GEOMETRIC INDEX OF PRICES FOR FIVE SPECIFIED COMMODITIES FOR THE YEARS 1927, 1928, AND 1929*

Commodities	1926			1927		1928		1929	
	Prices p_0	Productions q_0	Weights p_0q_0	Prices p_1	$p_0q_0 \log \frac{p_1}{p_0}$	Prices p_2	$p_0q_0 \log \frac{p_2}{p_0}$	Prices p_3	$p_0q_0 \log \frac{p_3}{p_0}$
Wheat	1.496	833 5	1246 9	1.372	− 46 8834	1 325	− 65.5869	1.180	−128.3060
Hogs	12.336	157.9	1947 9	10 137	−165 7663	9 628	−209 0097	10 324	−150 5727
Cotton[a]	0 175	8990 0	1573 2	0 176	4 0903	0 200	91.2456	0 191	59 4670
Petroleum	1 884	770 9	1452 4	1 285	−241 3889	1 203	−283.5085	1 233	−267 8226
Pig iron	18 548	39 07	724 7	17 697	− 14 8564	16 664	− 33 4811	18 189	− 6 0150
Total	. .	.. . .	6945 1	. .	−464 8047	.. .	−500 3406	.....	−493.2493
Weighted index	...	.. ...	100.0	. . .	85.7	.....	84.7	.....	84.9

* The weights are proportional to the values in exchange in 1926. Index for 1926 = 100.
[a] Price and production in pounds.

is the type of average to be used, the price relative (91.7) for wheat is multiplied by its value weight, the price relative for hogs (82.2), by its value weight, and similarly for all the commodities. The resulting products are added, and their total is divided by the sum of the weights. The quotient is the index number. If another type of average, as the harmonic average or the geometric average, is desired, the modifications in computation indicated by Tables 107 and 108 give the corresponding weighted index. The results, for the three types of average and for each year covered by Table 106, are shown in Table 109. In computing the weighted indexes

TABLE 109

INDEX NUMBERS OF PRICES FOR THE FIVE COMMODITIES, LISTED IN TABLE 104, WEIGHTED ACCORDING TO THEIR VALUES IN EXCHANGE, 1926*

Type	Method	1926	1927	1928	1929
Arithmetic	Fixed-base	100	86 5	86 4	86 3
	Link		86 5	99 1	100 5
	Chain	100	86 5	85 7	86 1
Harmonic	Fixed-base	100	84 9	83 1	83 6
	Link		84 9	98 6	100 0
	Chain	100	84.9	83.7	83 7
Geometric	Fixed-base[a]	100	85 7	84.7	84.9
	Link		85.7	98.8	100.2
Aggregate	Fixed-base[a]	100	86 5	86 4	86 3
	Link		86.5	99 9	99 8

* Unit: per cent of 1926 average.

[a] In the geometric and the aggregative, the chain index numbers are identical with the fixed-base numbers. Note also that the aggregative indexes are identical with the fixed-base arithmetic indexes.

of price relatives the weights are *values* in the base year. In the calculation of the weighted aggregate index number, on the other hand, the weights are the *physical quantities* exchanged in the base year. If an index number is computed from price relatives, the weights used must be value weights or weights proportional to these values; if based on actual prices, the weights must be actual quantities or weights proportional to these quantities. While it is possible to derive a median or a mode from a series of weighted relatives, the significance of the result is less apparent than in the three cases discussed, and indexes based upon weighted medians or modes are not likely to be encountered in practice.

In the calculation of the simple averages the arithmetic was found greater than the geometric, which in turn was greater than

the harmonic. Study of these weighted averages in Table 109 indicates the same differences—the arithmetic was found greater than the geometric, and the geometric greater than the harmonic. The purpose for which an index number is to be constructed determines the type of average and the weights to be used, and it is natural to assume that differences will occur depending upon the purpose of the index numbers constructed.

TYPE BIAS

Let it be assumed, for the present, that the purpose of an index number is to measure changes in the purchasing power of money and that the comparison of price levels is between two years only, the base year and a single given year. Certain requisites of a satisfactory index-number formula and certain desirable refinements in the selection of weights can then be brought out.[1] One requisite of a good index is that it satisfy the time-reversal test, which requires that the product of the index for the given year on the base year by the index for the base year on the given year shall be unity. This property obviously holds for the price relatives of a single commodity: the product of the relative for the given year on the base year p_1/p_0 by the relative for the base year on the given year p_0/p_1 is necessarily unity. It is also true, as appears from the algebraic formula, of the simple geometric index of price relatives. It is not true of the mean or harmonic types of indexes, and they are therefore said to have *type bias*. As the simple arithmetic mean is larger than the geometric mean, the former is said to have an *upward* type bias, and the harmonic mean similarly has a *downward* type bias.

One of the most effective ways of eliminating bias is by *crossing* index numbers having opposite bias. The crossing operation consists merely in taking an appropriate average: if two index numbers show the relation of the prices in a particular given year to the prices in the base year, and if one of these indexes has an upward bias and the other an approximately equal downward bias, the average of the two may be expected to yield an index which has less bias than either. The type of average to use in making the cross is determined by logical considerations. If the bias which is to be averaged out of the two given indexes is such that the absolute biases in the two indexes are about equal in magnitude

[1] Cf. IRVING FISHER, "The Making of Index Numbers," Houghton Mifflin Company, Boston, 1927, p. 62.

but different in sign, the arithmetic cross should be used; if the relative biases in two numbers are equal in magnitude but in opposite directions, the geometric average of the two will yield the most effective cross.

The upward bias of the arithmetic index is approximately of the same relative size as the downward bias of the harmonic, and therefore a geometric cross is desirable in this case. In fact, the geometric average of the simple arithmetic and harmonic indexes is algebraically identical with the simple geometric index, of the same pair of price relatives. The elimination of bias is therefore perfect in this case; and although elimination is less complete when there are more than two commodities, it is moderately effective. If excellent reasons for weighting index numbers did not exist, the discussion might end with the conclusion that the geometric index is preferable to the others.

WEIGHT BIAS; THE JOINT EFFECT OF THIS AND TYPE BIAS

As soon as weights are applied in the construction of an index, another kind of bias, called *weight bias*, appears. This bias is also called *weight correlation bias*, since it is caused by the correlation which is present between the weights and the price relatives.[1] Thus, usually, there is an upward weight bias when given-year values are used as weights, and a downward weight bias when base-year values are used. Therefore an index calculated with the simple arithmetic average or the simple harmonic average has a type bias, an index calculated with the weighted geometric average has a weight bias, and an index calculated with the weighted arithmetic average or the weighted harmonic average has both type bias and weight bias. (The term *weighted* is used here in a limited sense, meaning weighted with base-year values, or given-year values.)

Barring the simple geometric index, all indexes calculated by the formulas discussed above are affected with bias; each gives a result in comparing the given year with the base year, which is either too large or too small. The magnitude of the error thus introduced depends upon the nature of the individual price changes taking place between the two years: ultimately it is a question primarily of the dispersion of the series of price relatives. In a period in which price changes are small and fairly uniform

[1] Cf. W. M. PERSONS, "The Construction of Index Numbers," Houghton Mifflin Company, Boston, 1928, p. 10; also EDWIN FRICKEY, "The Theory of Index-Number Bias," *Review of Economic Statistics*, November, 1937, pp. 161–173.

over the whole list of commodities, the dispersion and skewness coefficients are not large and the effects of bias are slight; but in a period in which changes in the price level of certain commodities relative to the general level are great, the errors may be considerable.

The fact that the type bias of the arithmetic index is upward and the weight bias due to base-year weighting is downward, does not mean that the net bias is zero. Theoretically, the type bias in one direction is not offset by the weight bias in the opposite direction. It does suggest the possibility, however, that the net bias of the weighted arithmetic index with base-year values as weights may be small. More precisely, the net bias of an index constructed with this formula is not likely to be large, except at times when the movement of prices is very erratic. As several of the leading indexes rest upon this formula, the practical importance of this partial elimination of bias is evident.

In like manner the weighted harmonic index, with given-year values as weights, tends to have small net bias, since its type bias and weight bias are in opposite directions. As the net bias is likely to be in opposite directions and of about equal relative magnitude for the arithmetic index with base-year weights and the harmonic index with given-year weights, the crossing of these two indexes should yield an index almost free from bias.

PROFESSOR FISHER'S IDEAL INDEX

The result of this operation (using a geometric cross) is Professor Fisher's Ideal Formula

$$I = \sqrt{\frac{\Sigma p_i q_o}{\Sigma p_o q_o} \frac{\Sigma p_i q_i}{\Sigma p_o q_i}}$$

where the subscript i refers to the given year and o refers to the base year, and p signifies the price of a particular commodity in the index, and q the quantity of that commodity exchanged.[1]

A more fundamental form of the Ideal Formula is

$$I = \sqrt{\frac{\Sigma p_o q_o \frac{p_i}{p_o}}{\Sigma p_o q_o} \frac{\Sigma p_i q_i}{\Sigma p_i q_i \frac{p_o}{p_i}}}$$

[1] For full discussion, see FISHER, *op. cit.*, p. 220.

and the factors in this expression reduce to those shown above for the following reasons: When the relative p_1/p_0 is multiplied by the base value p_0q_0 the result is obviously p_1q_0. This is clearly the value of the base-year quantity at the given-year price. That is, in multiplying the relative by the weight in the case of the weighted mean, we are in effect calculating the value which the quantity exchanged in the base year would have at the average price of the given year. On the other hand, the sum of the weights, used in the denominator of the weighted arithmetic average, is merely the total value in exchange in the base year. Hence, the net effect of calculating the weighted arithmetic index number is to get the ratio of the total value of a specific bill of goods (those exchanged in the base year) at the average prices of the given year to the total value of the same bill of goods at the average prices of the base year. In other words, the average of price relatives, weighted with the base values as weights, is the same as the ratio of the weighted aggregate of given-year prices to the weighted aggregate of base-year prices, with the base quantities as weights. An analogous algebraic relation holds for the harmonic index weighted with given-year values, and accounts for the modification of the second factor in the above formula.

FURTHER TESTS OF INDEXES

A second important test, applicable to weighted index numbers, is the *factor-reversal test*, which requires that the formula shall serve equally as well for an index of changes in prices as for an index of changes in quantities, it being understood that to get the quantity index the factors p and q in the formula for the price index are merely interchanged. Then, that the test be satisfied, the product of the price index by the quantity index must equal the ratio of the aggregate value in the given year to the aggregate value in the base year. This equality obviously holds when the price relative and the quantity relative for a single commodity are multiplied together, and the test merely specifies that this same relation must hold for the index numbers, of prices and of quantities, of a group of commodities.

Neither the simple nor the weighted form of any of the elementary index numbers—arithmetic, harmonic, geometric—satisfies the factor-reversal test. The Ideal Formula does satisfy the test, and this is one of its advantages as a scientific measuring device. The requirement imposed by the factor-reversal test is, however,

narrowly specialized; and it is largely ignored in the use of many of the less complicated formulas in actual practice.

A third test, applicable to index numbers involving comparisons between more than two years (dates), is the *circular test*, which requires that a formula be such that the index number it yields for a particular year referred to the base year shall be the same whether it be computed by the fixed-base or by the chain process. This, again, is a property possessed by the price relatives of a single commodity. In the case of the circular test, the geometric index, both in its simple form and when weighted with fixed values, has the advantage; it satisfies the test, whereas the arithmetic and the harmonic do not. The aggregate with fixed weights also satisfies the circular test.

CONSTANT WEIGHTS MORE PRACTICAL THAN VARIABLE WEIGHTS

If the weighted geometric index satisfied the time-reversal test, the evidence would strongly favor the systematic use of the geometric formula for index numbers. Its failure to satisfy the time-reversal test, which is another way of describing its liability to a weight bias, is an important objection against it for all problems requiring a very precise measurement of changes from year to year. If the utmost precision in the formula is desired, if the type of average employed is to meet the considerations discussed above, one of the more complicated index numbers must be used. Fisher's Ideal Formula is one of these for a comparison of two periods only, and another, which is nearly as satisfactory and requires less extensive calculations, is

$$I = \frac{\Sigma p_1(q_0 + q_1)}{\Sigma p_0(q_0 + q_1)}$$

The Ideal Formula only approximately satisfies the circular test; but as this formula is especially adapted to the comparison of two specific years rather than a series of several years, the circular test is less important for it. The great obstacles to the use of the formula are that the variable weighting necessitates rather laborious computation and renders the significance of the comparison measured by the index difficult of comprehension. This obscurity arises because the changes from year to year, as indicated by an index with variable weights, may be due not entirely to changes in prices, but also to changes in the quantities.

For the great bulk of index-number analysis sufficient accuracy is afforded by the arithmetic average of relatives weighted with a

fixed set of values, or its equivalent, the ratio of two aggregates of prices each weighted with a fixed set of quantities. One of the standard general-purpose wholesale-price indexes in the United States, the index of the United States Bureau of Labor Statistics, is of this sort. It is at present the ratio of the weighted aggregate of prices in the given year (or month) to the weighted aggregate of the average prices of 1926; the weights are the quantities entered in trade in the period 1933–1935.

CLASS AND COMMODITY WEIGHTS

A special problem frequently arises in the construction of weighted index numbers, leading to the development of systems of commodity weights and class weights. For any one group of commodities, say those of any industry, the value weights of the individual commodities within the group can be estimated according to the principles outlined above. By the use of these commodity weights, an index can be obtained for the group as a whole. Likewise, for any other one group, the commodity weights can be determined and an index calculated for the group. When it comes to the computation of a general index for "all commodities," it may not be fitting to combine in one index the price relatives for all the commodities, irrespective of grouping, by the use of their commodity weights. The reason is that the quotations in one group, say the foods, may be much more comprehensive than those in another, say the metals. Hence, if all commodities are combined by the use of the commodity weights, undue weight is given to the commodities in the food group. The difficulty can be met by using class weights, which give merely the relative weights of the several groups. Thus, although the data for metals are less adequate than those for foods, it may be possible to estimate with fair accuracy the relative importance of the foods group and metals group in the price structure as a whole. Then to get the index number for "all commodities" the procedure is to construct an average of the group indexes with their respective class weights. The same observations apply to the construction of a group index from several subgroup indexes, each derived from commodity data. The data of Table 110 illustrate the difficulty, although the discrepancies between the group representations are not so striking in this instance as in some practical cases. The ratio of the estimated values of a particular group as a whole to the aggregate of the commodity weights (values in exchange) of the commodities actually used in getting a price index for that group is

called the *class weight.* In deriving the index for the major group, which comprises the several groups, the individual group indexes are weighted with these ratios. If it is not desired to know the individual group indexes, the commodity weights can be "corrected" by use of the class weights, and the general index can then be computed directly from the commodity prices and the corrected weights.

TABLE 110
DERIVATION OF CLASS WEIGHTS*

Group designation	Estimated value of group* (1)	Total of the commodity weights* (2)	Class weight (1)/(2)
Food Group	17,514	37,498	
Feed and forage	1,859	8,619	0 216
Wheat and wheat products	2,208	3,883	0 569
Corn and corn products	1,046	5,355	0 195
Oats, rice, buckwheat, and their products	112	151	0 742
Barley, hops, rye, and their products	881	907	0 971
Sugar and related products	783	1,041	0 752
Vegetables and truck	582	1,213	0 480
Edible vegetable oils	282	949	0 297
Fruits, nuts, and wine	361	674	0 536
Spices and condiments	16	14	1 143
Tea, coffee, and cocoa	200	127	1 175
Tobacco and tobacco products	1,200	291	4 124
Livestock, meats, and fats	4,446	9,337	0 476
Poultry and dairy products	3,379	4,752	0 711
Fish and oysters	159	185	0.869

* Unit million dollars. Source. W. C. Mitchell, "History of Prices during the War," Washington, War Industries Board (Price Bulletin 1), 1919, p. 24.

STATISTICAL DEFLATION

Many statistical series pertaining to economic factors are presented in terms of value. Bank debits, for example, represent the dollar volume of all transactions by bank check. Building activity is measured by the dollar volume of contracts awarded or by the value of building permits. Total imports and total exports are stated in terms of dollars. Fluctuations in these value series over a period of time are caused not only by changes in physical quantities but also by changes in prices. The elimination of that part of the fluctuation caused by price movements from these value series would result in measures of physical quantities, and for many purposes in economic analysis study of changes in

physical quantities over a period of time is desirable. Such elimination of the price movement from a value series is called *statistical deflation;* it consists in dividing the value series over a period of time by the corresponding price indexes.

The success of the deflation process depends upon the construction of the proper index number of prices to be used as a deflator. The obtaining of the price index is exceedingly difficult, particularly because of the changes in weights which take place in both the long and the short run. An index number constructed with fixed weights is not suitable, for the deflation of a value series by such an index number would assume that the individual quantities that go to make up the quantity element in the total value for any period of time are always proportional to the fixed weights in the index number, an assumption that obviously will not always be true. The difficulty of obtaining the proper price index is the chief obstacle to the use of statistical deflation of value series to obtain fluctuations in physical volumes. If not used carefully, statistical deflation will result in fictitious fluctuations in the quantity data obtained.[1]

---

[1] Cf. W. L. CRUM, "Discussion: Deflated Dollar-Value Series as Measures of Business," *Review of Economic Statistics*, April, 1926, pp. 92–100.

CHAPTER XX

SECULAR TREND

FOUR COMPONENTS OF TIME FLUCTUATION

The method of determining relatives and index numbers discussed in the two preceding chapters furnishes a means of measuring the fluctuations in an economic factor. The present and subsequent chapters develop devices for analyzing the nature of such fluctuations in any single economic factor, and for comparing the fluctuations in several factors. In order to simplify the analysis and to clarify the statistical description, the total fluctuation in any single factor may be considered as resulting from the simultaneous action of several component fluctuations. Thus, if the index of the retail prices of eggs for December 14, 1937, compared with the 1923–1925 average, is 77, the indicated change of 23 per cent from the base period may be assumed to arise from the joint action of several causes. It is presumably quite impossible to separate each of these causes from the others and measure each alone with precision; but the causes can be classified into fairly distinct groups, and the group effects, each contributing its share of the total change of 23 per cent, can be measured. Thus a more detailed picture of the fluctuation is formed than that furnished by the index number alone.

The classification of causes can, of course, be carried out according to widely different plans, but one scheme of grouping, which conforms to a satisfactory economic description of the fluctuation and is well adapted to statistical treatment, divides the fluctuation into four parts. These parts are: secular trend, seasonal variation, cyclical fluctuation, and irregular deviations. The *secular trend* is that part of the fluctuation which is due to the gradual and persistent tendency to change which exists for an interval of several years, an interval the definite length of which in general cannot be assigned but which may extend over a generation and sometimes for a longer period. *Seasonal variation* is that part of the fluctuation which recurs annually with the seasons, by which a particular month of the year tends always to deviate in a certain direction and to a certain degree from the

normal for the year. *Cyclical fluctuation* is the approximately rhythmical sweep from high to low and back to high, reaching usually over an interval of three or four years, which marks the ebb and flow of business prosperity and depression. Finally, there are occasional abrupt and extreme deviations, usually explicable as marking the beginning of war or the outbreak of financial panic or the occurrence of some other strikingly exceptional economic event; these deviations appear to have no uniform relation to the ordinary smooth course of economic movement and are therefore designated *irregular deviations.*

In the illustration considered above, the part which might be played by each of the four fluctuation groups is apparent. A tendency, persisting over a long interval of years, for the purchasing power of money to rise (for the general level of prices to fall), progress in agricultural methods, and gradual changes in marketing and transportation would contribute some portion of the 23 per cent change coming under the head of secular trend. The fact that December is a winter month would undoubtedly result in a retail price of eggs considerably above that which holds on the average throughout the year: this discrepancy is the effect of seasonal variation. The fact that December, 1937, was a month in a period of depression, which, however, was less severe than that which was at its worst in 1932 and 1933, suggests that the retail price of eggs was still much below normal because of the cyclical fluctuation. If December, 1937, had been marked by the outbreak of a severe railroad strike which interrupted the transportation of eggs to market, there would have been an extraordinary advance in price and the price relative 77 might easily have been 100 or even much higher: the increase due to the strike would have been an irregular deviation.

As each of the four types of fluctuation may be manifest in different degrees in two different series, a satisfactory comparison of such series involves a separate comparison for each type of fluctuation. For example, direct comparison of the series of price indexes for eggs and a series showing factory employment may be quite futile, because of the different trends or the different seasonal movements of the two series. Interest generally centers, however, in a comparison of the cyclical movements in two series, although the significance of other comparisons between economic series is considerable and is now receiving greater consideration than formerly. In any case, the total fluctuation of each series needs separation into the component types, in order to afford a scientific

basis of comparison. This process of analysis consists of the measurement of each type of fluctuation in the given series: strictly speaking, certain types are measured independently, and the elimination of these independent types from the original time series gives us the remaining types of fluctuation.

On the whole, a statistical device for eliminating the effect of irregular deviations cannot be developed, although their presence is usually obvious, and they can ordinarily be examined in the light of nonstatistical considerations. Hence, in a statistical analysis the irregular deviations are customarily grouped with the cyclical fluctuation and this composite is regarded as the cyclical fluctuation. Otherwise the statistical analysis follows the plan outlined above: the total fluctuation is classified by arithmetical operations into secular trend, seasonal variation, and cyclical fluctuation, and each of these three major influences is considered separately. The remainder of this chapter will be devoted to a discussion of the measurement and elimination of secular trend.

While the treatment of the measurement of secular trend in the following pages is primarily statistical in character, the student should realize that elements of a nonstatistical nature are very important in ascertaining the true secular trend. Attention must be paid to the fundamental forces which cause the trend to move in a certain way and at a certain rate if the trend derived through statistical processes is to be significant. Historical analysis and a knowledge of economic theory also aid in a proper understanding of the problems involved in measuring the secular trend.

GRAPHIC EXAMINATION OF THE TREND

The most satisfactory approach to the problem of trend for any particular series is by examination of the chart displaying the actual items. Table 111 gives the total production of boots, shoes, and slippers, monthly, November, 1921, to February, 1938, and Chart 104 shows the same data, with an arithmetic vertical scale. A brief inspection of this chart is sufficient to disclose the characteristic movements of this production series; within each year the seasonal movement can be noted—production ordinarily increases during the first quarter, declines in the second quarter, reaches a peak in the third quarter, and declines in the last quarter to the low point of the year in December; also noteworthy is a succession of wavelike fluctuations and the persistence through these fluctuations of a strong upward drift in production. The fluctuations, with their low points in 1924, 1930–1933, and late 1937, are the

TABLE 111

TOTAL PRODUCTION OF BOOTS, SHOES, AND SLIPPERS IN THE UNITED STATES, MONTHLY, NOVEMBER, 1921–FEBRUARY, 1938*

Month	1921	1922	1923	1924	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938
January		25 12	30 74	26.50	26 08	23 87	24 99	26 21	27.25	26 53	19.89	21 23	22.72	26.04	29 56	33 36	37.15	25.52
February		24 55	30 30	26 83	26 45	25 70	27 29	29 63	27 71	25 90	23 97	25 96	26 38	30 53	30 87	33 05	39.58	29.77
March		29 35	35 84	28 86	29 89	29 93	31 28	32 30	30 90	28 63	29 36	30 68	28 58	35 55	34 23	34 83	46 12	
April		26 85	31 87	28 00	29 48	26 64	28 39	26 63	29.38	29 00	29 89	25 95	27 63	34 42	34 56	33 40	40.30	
May		26 23	30 93	25 24	25 11	23 13	25 63	26 43	29 16	24 51	28 45	22 50	32 97	34 06	31 26	30.26	35.41	
June		24 82	28 27	22 46	23 45	25 04	27 50	27 28	28 12	23 90	27 84	23.56	34 86	28 54	27.23	29.37	34.45	
July		22 69	25 26	21 39	24 76	25 05	27 78	28.15	30 22	24.12	28.61	20 44	33 75	28 39	32.27	35.68	34 84	
August		27 68	30 03	25 47	28 49	29 65	35 06	34 97	36 44	28 43	33 47	30 78	37 02	35 62	37 24	40.67	38 66	
September		28 29	27 55	27 72	29 77	31 67	33 93	31 00	34 83	29 33	31.29	33 88	31 23	28 18	33 91	40 97	34 03	
October		30 37	30 70	30 83	31 06	31 66	32 27	33 39	37 19	27 73	25 38	33 07	31 46	28 71	35 95	39 92	29.09	
November	23 53	30 08	26 95	25 32	24 63	26 76	25 97	26 44	27 72	18 54	18 52	25 15	23 69	23 85	27 71	30.34	21.29	
December	24 13	27 85	22 68	24 60	24 40	25 42	23 52	21 91	22 48	17 54	19 56	20 10	20 09	23 20	28.95	33 38	21.05	

* Unit: million pairs. Source: Monthly reports of the U S. Department of Commerce, Bureau of the Census. (Rubber-soled footwear is excluded.)

cyclical swings arising from alternating prosperity and depression. The steady upward drift which, with passing years, raises the average level about which the fluctuations take place in the secular trend, may be represented approximately by a straight line that extends upward to the right.

With such a chart available one can represent the trend by stretching a string (an elastic string is desirable in practice) across the face of the chart. One adjusts the string, by changing its inclination and by shifting it parallel to itself, until it "fits" the curve as closely as possible. A crude test of the fit requires

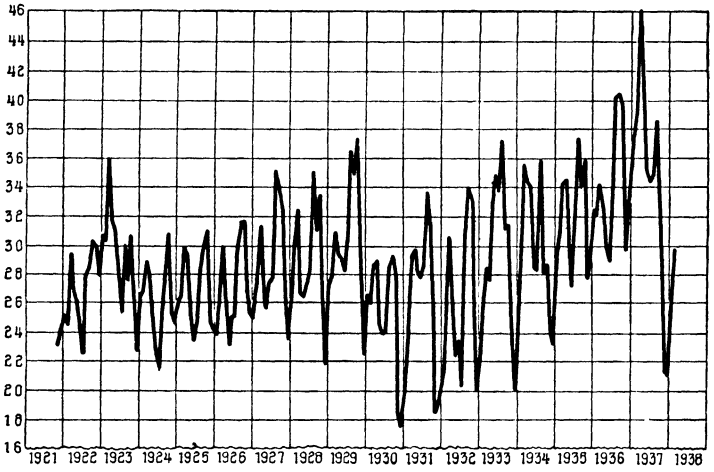


CHART 104.—Total production of boots, shoes, and slippers in the United States, monthly.
(Unit: million pairs. Data in Table 111)

that each pair of adjacent high and low points of the curve, for example those of 1923 and 1924, fall about the same distance above and below the line. This requirement cannot, and need not, be met consistently throughout the chart: it is more important to consider the balance about the line of a whole section of the curve lying above the line with the adjacent section below the line (for example, the section in 1928 and 1929 with that of 1930–1932). It should not be expected that even this balancing of sections will be perfect, but the more nearly it is realized the better the general fit of the line.

The balancing cannot ordinarily, moreover, be equally well realized in the several cycles of a long series. Trial will show that

the position of the line differs as different segments of the interval 1922–1937 are examined. If the string is applied to different portions of the interval 1922–1937, different results are obtained,

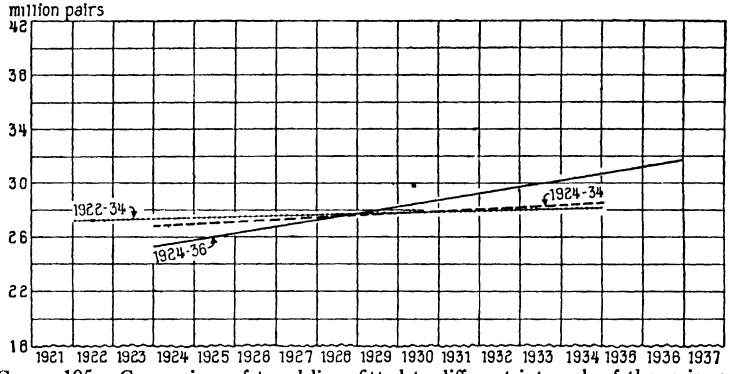


CHART 105.—Comparison of trend lines fitted to different intervals of the series of Chart 104.

and several of these appear in Chart 105. The differences in trend lines when based on different intervals are not always so marked, however, as they are in this chart. In some cases the distinctions (Chart 106) may be so fine that visual fitting of the

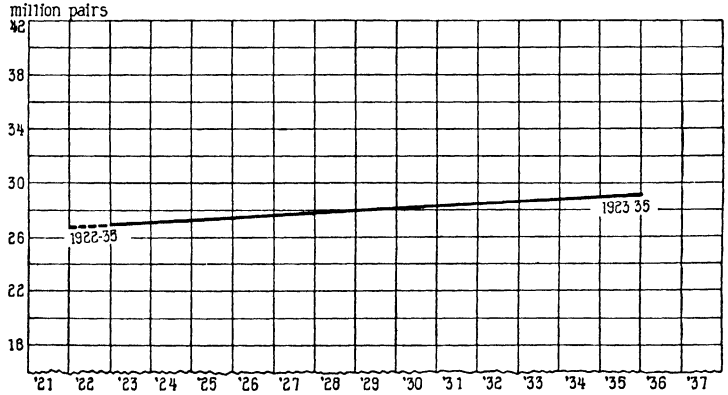


CHART 106.—Trend lines which belong to quite different intervals, but which are not markedly different.

(Based on data of Table 111, and Chart 104.)

string does not reveal them, but the method presented below for numerically measuring the trends will bring out even the less prominent differences arising from changes in the trend interval.

The actual selection of the interval used in fitting the trend is ordinarily made by inspection of the original curve (Chart 104). The minimum length of the interval should be several (two or more) complete cycle waves; an interval 1924–1929 would be too short. If this requirement is ignored, the "trend" as computed, by the method given below, is likely to be erroneous in that it will be determined partly by the cyclical fluctuation as well as by the underlying trend. In fact, one of the essential considerations governing the selection of the interval is that the effect of the cyclical movements, in their influence upon the *apparent* trend, should be reduced to a minimum. It should be remembered that secular trend is due to the gradual and persistent tendency for data to change over a *long* period of time, and for this reason an accurate measure of trend is possible only when based on a relatively long period—at least ten years.

The choice of beginning and terminal years in the interval generally is on the following principle. The cyclical movements at the beginning and at the end of the interval should be in symmetrical phases: if there is a rise at the beginning, there should be a decline at the end (1922 and 1930 or 1931); if there is a peak at the beginning, there should be a peak at the end (1923 and 1929 or 1936). Moreover, if possible, the first and last years should be chosen so that the cyclical movement is in an intermediate phase rather than at the peak or trough: 1926 to 1930 or 1931 is better than 1923 to 1929 or 1924 to 1932. If the terminals are thus in intermediate phases, however, the *level* of the line—but not its direction—may be somewhat in error in cases in which the interval is short.

The rules laid down above for the selection of the beginning and ending years of the interval are less important if the interval selected is comparatively long. The period 1923–1936, in which the beginning and terminal years of the interval are at cyclical peaks, is probably the best interval to use for the computation of the secular trend in this case. For reasons of exposition, however, the somewhat longer period 1922–1936 is selected. The differences between a trend based on this interval and one based on the interval 1923–1936 are not great. Further attention will be given below to the problem of selecting the interval, and another decision which must be made before computation begins. In the present case, however, the visual evidence is fairly clear: a single straight line appears to "fit" very well for the interval 1922–1936.

The technique of measuring and eliminating the trend numerically will, therefore, be developed from the data in Table 111.

DETERMINATION OF THE LINE OF TREND BY COMPUTATION

The standard method of determining the position of the line of trend is an application of the principle of least squares. The *line of trend* is defined, on this principle, as a line located on the chart so that the sum of the squares of the vertical deviations of the points of the given curve from the line of trend is as small as possible. Thus, in Chart 107, AB is the trend line if the sum of the

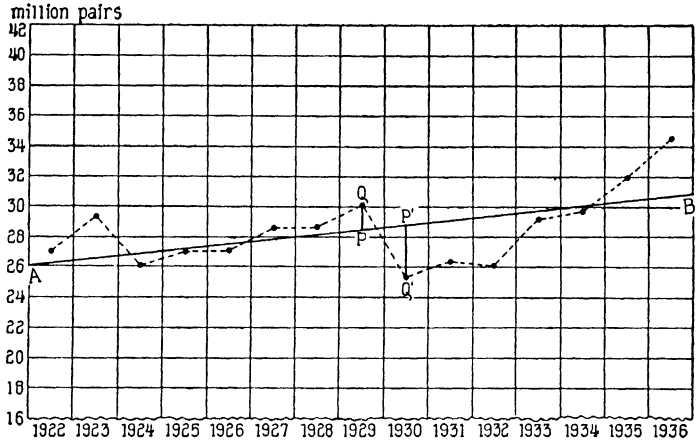


CHART 107.—Illustration of the principle of location of the line of trend.

squares of PQ, P^1Q^1 . . . etc. is a minimum. By this definition, the line of trend is obviously the same as the line of regression through those points (Chap. XVI): the line of trend for production of boots, shoes, and slippers is "the line of regression of production on time." Some other line, of course might be used to represent the secular trend, but the line determined by the principle of least squares, the "average" line, is widely considered as the one giving the best fit to the actual data.

In practice, lines of trend are generally fitted to annual figures (as in Chart 107) rather than monthly figures. Annual averages of the monthly figures are found, and the trend is calculated from these annual average items. The use of annual figures largely eliminates the influence of seasonal variation from the determination of trend, and it can be shown algebraically that the line fitted

to the annual items does not differ substantially from that fitted to the monthly items. The monthly averages for each year for the data of Table 111, 1922–1936, appear in column 2 of Table 112, and Chart 107 shows the curve of these annual items.

TABLE 112

COMPUTATION OF THE LINEAR TREND FOR THE PRODUCTION OF BOOTS, SHOES, AND SLIPPERS IN THE UNITED STATES, 1922–1936*

Year	Monthly average production Y	x	Yx	
			−	+
(1)	(2)	(3)	(4)	(5)
1922	26.99	−7	188 93	
1923	29.26	−6	175 56	
1924	26 10	−5	130 50	
1925	26 96	−4	107 84	
1926	27 04	−3	8i 12	
1927	28 63	−2	57.26	
1928	28 70	−1	28.70	
1929	30 12	0		
1930	25 35	+1		25 35
1931	26 35	+2		52.70
1932	26 11	+3		78 33
1933	29 20	+4		116 80
1934	29 76	+5		148 80
1935	31 98	+6		191 88
1936	34 60	+7		242 20
Total	427 15		−769 91	+856 06

Net total = + 86.15

$a = \frac{427.15}{15} = 28\,477.$ $b = \frac{86.15}{280} = 0\,3077$ for a period of 1 year

* Unit: million pairs.

The fact that the trend line is the line of regression of production on time points the way to the numerical treatment of the problem. If the two variables, production and time, are represented by Y and X, the problem is to find the regression equation of Y on X.[1]

[1] The regression equation is (Chap. XVI)

$$Y - M_y = r\frac{\sigma_y}{\sigma_x}(X - M_x)$$

where M_y and M_x are the means and σ_y and σ_x are the standard deviations for Y

This relation is often written

$$Y = a + bx$$

where

$$b = \frac{\Sigma Yx}{\Sigma x^2}, \qquad a = M_y = \frac{\Sigma Y}{N}$$

and where x is measured from the center of the time period used in determining the trend.

The two numbers, a and b, determine the position of the line of trend: a is called the *central ordinate* and b is called the *increment rate*. The central ordinate is the size of Y for the instant when x is zero, that is, the ordinate at the center of the interval. The increment rate is the amount of change in Y for a unit change in x.

The details of the computation of trend for the production series under study appear in Table 112. The items of column 3 are the values of x: the origin of time is chosen at the middle of the interval selected for fitting the trend, July 1, 1929; and the unit is taken as one year, in order that the time distances to the centers of the calendar years—the annual item is assumed fixed at the center of the year—may appear as whole numbers. Columns 4 and 5 give the products Yx, and the net total of these columns is the numerator of b. The denominator of b is obtained by summing the squares of the items of column 3, or may be read from Table 113, which gives Σx^2 for intervals of different lengths. The total of column 2 divided by the number of years in the interval yields a. The value of a is ordinarily calculated to one more

---

and X, and r is the correlation coefficient given by

$$r = \frac{\frac{\Sigma YX}{N} - M_y M_x}{\sigma_y \sigma_x}$$

In applying the method, the origin of time is usually taken at the center of the interval, and the time deviations from this point are represented by x. Evidently this renders $M_x = 0$, and r becomes

$$r = \frac{\Sigma Yx}{N\sigma_y \sigma_x}$$

and hence the line of trend (regression) is

$$Y - M_y = \frac{\Sigma Yx}{N\sigma_x^2} x$$

which, by definition of σ_x^2, reduces to

$$Y = \frac{\Sigma Yx}{\Sigma x^2} x + M_y$$

decimal place than the original data and the value of b is usually calculated to two more decimal places than the original items, so that the trend values, when expressed in the same number of decimal places as the original data, will not be subject to any rounding errors.

TABLE 113

THE DENOMINATOR OF THE INCREMENT RATE, b, FOR INTERVALS OF VARIOUS LENGTHS

Length of interval (years)	Σx^2, for number of years		Length of interval (years)	Σx^2, for number of years	
	Odd	Even*		Odd	Even*
6		70	19	570	
7	28		20		2,660
8		168	21	770	
9	60		22		3,542
10		330	23	1,012	
11	110		24		4,600
12		572	25	1,300	
13	182		26		5,850
14		910	27	1,638	
15	280		28		7,308
16		1,360	29	2,030	
17	408		30		8,990
18		1,938			

* The summation of half-year intervals.

The values of a and b are substituted in the equation

$$Y = a + bx$$

to yield the equation of the line of trend. The equation of the line of trend for annual average production of boots, shoes, and slippers based on the period 1922–1936 is then

$$Y = 28.477 + 0.3077x$$

where Y represents annual average production expressed in units of million pairs and x represents the number of years measured from July 1, 1929. From this equation the trend value of any particular year can be determined by substituting the appropriate value of x in the equation.

The computation of the trend equation can be abridged by pairing the actual items according to their distances from the center of the interval. Thus in Table 114 the item for 1922 is

seven years before the center and that for 1936 is seven years after the center. In one case x is negative and in the other positive; and consequently, in the computation of the product terms, the number 7 can be multiplied by the difference between the two items. A similar pairing is made of the other items equally distant from the center.

This short method of computation brings out an elementary test of the goodness of fit of the line of trend. If there were no cyclical fluctuations to disturb the regularity of the growth, the

TABLE 114

ABRIDGED COMPUTATION OF THE LINEAR TREND FOR THE PRODUCTION OF BOOTS, SHOES, AND SLIPPERS IN THE UNITED STATES, 1922–1936*

Year	Y_{-x} (1)	Year	Y_x (2)	$Y_x - Y_{-x}$ (3)	x (4)	$x(Y_x - Y_{-x})$ (5)	$\frac{Y_x - Y_{-x}}{x}$ (6)
1929	30 12	. .	. .	.	0	.	
1928	28 70	1930	25 35	−3 35	1	− 3 35	−3 35
1927	28 63	1931	26 35	−2 28	2	− 4 56	−1 14
1926	27 04	1932	26 11	−0 93	3	− 2 79	−0 31
1925	26 96	1933	29 20	+2 24	4	+ 8.96	+0.56
1924	26 10	1934	29 76	+3 66	5	+18 30	+0 73
1923	29 26	1935	31 98	+2 72	6	+16.32	+0.45
1922	26 99	1936	34 60	+7 61	7	+53 27	+1.09
	223.80		203 35			+96 85	
			223 80			−10 70	
			427.15			+86.15	

$$a = \frac{427.15}{15} = 28.477 \qquad b = \frac{86.15}{280} = 0.3077$$

* Unit: million pairs.

only change from year to year would be that due to trend. Hence, for a straight-line trend, the differences shown in column 3 of Table 114 should be proportional to x as given in column 4. In other words, the ratios of the items of column 3 to the corresponding values of x should be substantially equal to each other. In so far as they vary from equality, there is lack of perfect fit (see column 6).

Although the principle of location of the line of trend and the plan of computation are unchanged, the actual details differ slightly if the number of years in the interval is even. To illustrate this case, a line of trend is fitted to the production series for the interval 1923–1936. Tables 115 and 116 show the calculations by

both the long and abridged methods, and here x is measured in a different unit: in terms of half years from the center (January 1, 1930 of the interval 1923–1936. The mid-point of 1929 is one half year before January 1, 1930, and that of 1930 one half year after the center of the interval; the mid-point of 1923 is thirteen half years before January 1, 1930, and that of 1936 is thirteen half

TABLE 115

COMPUTATION OF THE LINEAR TREND FOR THE PRODUCTION OF BOOTS, SHOES, AND SLIPPERS IN THE UNITED STATES, 1923–1936*

Year	Monthly average production Y	x	Yx	
			−	+
(1)	(2)	(3)	(4)	(5)
1923	29 26	−13	− 380.38	
1924	26.10	−11	− 287.10	
1925	26 96	− 9	− 242 64	
1926	27.04	− 7	− 189.28	
1927	28 63	− 5	− 143.15	
1928	28 70	− 3	− 86.10	
1929	30 12	− 1	− 30.12	
1930	25 35	+ 1		+ 25 35
1931	26 35	+ 3		+ 79 05
1932	26 11	+ 5		+ 130.55
1933	29 20	+ 7		+ 204 40
1934	29 76	+ 9		+ 267.84
1935	31 98	+11		+ 351.78
1936	34 60	+13		+ 449 80
Total	400.16		−1358.77	+1508.77

Net total = + 150.00

$a = \frac{400.16}{14} = 28.583$ $\quad b = \frac{150}{910} = 0.1648$ for a period of one half year

* Unit: million pairs.

years after. The values of a and b differ slightly from those calculated on the interval 1922–1936. The equation of the line of trend based on the interval 1923–1936 is

$$Y = 28.583 + 0.1648x$$

where Y represents the annual average production expressed in units of million pairs, and x represents the number of half years measured from January 1, 1930. As in the preceding equation, the trend values of any particular year can be determined by substituting the appropriate value of x in this equation.

The difference in results, according as different intervals are used, is of basic importance. Whereas the actual calculation of the increment rate and central ordinate are strictly numerical operations depending upon the data alone, the selection of the interval rests upon the judgment of the investigator. Although he makes his decision in the light of evidence afforded by the chart, it is a decision about which different observers may reasonably differ. In some series little room for doubt about the interval exists, but in others two or even three selections may appear

TABLE 116

ABRIDGED COMPUTATION OF THE LINEAR TREND FOR THE PRODUCTION OF BOOTS, SHOES, AND SLIPPERS IN THE UNITED STATES, 1923–1936*

Year	Y_{-x} (1)	Year	Y_x (2)	$Y_x - Y_{-x}$ (3)	x (4)	$x(Y_x - Y_{-x})$ (5)	$\frac{Y_x - Y_{-x}}{x}$ (6)
1929	30 12	1930	25 35	−4 77	1	− 4 77	−4 77
1928	28 70	1931	26 35	−2 35	3	− 7 05	−0 78
1927	28 63	1932	26 11	−2 52	5	− 12 60	−0 50
1926	27 04	1933	29 20	+2.16	7	+ 15 12	+0 31
1925	26 96	1934	29 76	+2.80	9	+ 25 20	+0 31
1924	26 10	1935	31 98	+5 88	11	+ 64 68	+0 53
1923	29 26	1936	34 60	+5 34	13	+ 69 42	+0 41
	196 81		203 35			+174 42	
			196 81			− 24 42	
			400 16			150 00	

$a = \frac{400.16}{14} = 28.583$ $\quad b = \frac{150.00}{910} = 0.1648$ for a period of one half year

* Unit: million pairs.

equally appropriate from an inspection of the chart. Fortunately, in cases of the latter sort, actual calculation generally shows that the trend lines are not seriously different for the different intervals. It is ordinarily safe, therefore, to accept a decision resting upon a reasonable interpretation of the chart of actual items, according to the rules enunciated above (page 304).

THE ORDINATES OF TREND

As has been pointed out, the calculated results for the central ordinate a and the increment rate b, when substituted in the equation $Y = a + bx$, yield the equation of the line of trend. From this equation the various ordinates of trend during the period can be derived. The *ordinate of trend* for any date is the value of Y

for the point on the line of trend at that date: it is the size which the variable would have had at that date if there had been no other fluctuation except the trend. In Chart 108, for instance, the ordinate for July 1, 1929, is the distance *RQ*.

The central ordinate is the ordinate of trend for the middle of the interval; it is that value of Y obtained when $x = 0$—for example, for July 1, 1929, in the 1922–1936 interval, and for

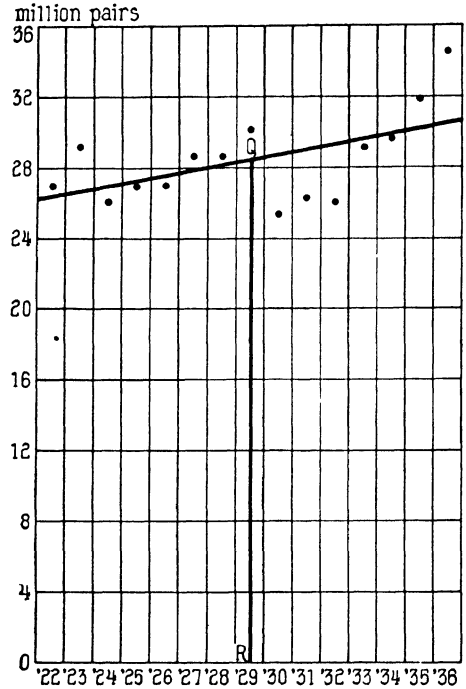


CHART 108.—Illustration of the position of the *central ordinate* of the line of trend.

January 1, 1930, in the 1923–1936 interval. The increment rate gives the change in the height of the ordinate of trend per unit of time—per year for an interval with an odd number of years (as 1922–1936), and per half year for an interval with an even number of years (as 1923–1936). In the first case the annual increment is b and the monthly increment $b/12$; in the second case the *annual* increment is $2b$ and the monthly increment $b/6$.

Annual ordinates of trend for the interval 1922–1936 are determined by substitution of the appropriate values of x in the trend equation.

$$Y = 28.477 + 0.3077x$$

where, it will be remembered, Y represents annual monthly average production of boots, shoes, and slippers expressed in units of million pairs and x represents the number of years, measured from July 1, 1929. The trend ordinate for 1929, when $x = 0$, is 28.48 million pairs. The trend value for 1930 is obtained by substituting for x in the equation of trend the number 1; that for 1931, by substituting $x = 2$, etc. Trend ordinates are easily obtained by the following method (Table 117):

Year	x	Equation	Ordinate
1927	−2	28 477 − 2(0.3077)	27.86
1928	−1	28 477 − 1(0.3077)	28 17
1929	0	28 477 + 0	28 48
1930	1	28 477 + 1(0 3077)	28 78
1931	2	28 477 + 2(0.3077)	29.09

These computations are most efficiently performed on a calculating machine.

When the trend equation is based on an interval in which the number of years is even, as is the case for the period 1923–1936, x is measured in half years from the center of the interval, and substitution of those values of x given in column 3 of Table 115 in the equation of trend will yield the trend ordinates for each of the years in the period 1923–1936.[1]

[1] It is always possible, of course, to change the equation of trend even when based on an even number of years so that the ordinates will refer to the mid-point of the year. The equation for the period 1923–1936 as computed previously is

$$Y = 28.583 + 0\ 1648x$$

where x is measured in half years from January 1, 1930. It is necessary to change this equation in such a way that the central ordinate will refer to July 1 rather than January 1 and also to change the increment (b) so that it refers to the change which takes place in one year rather than in one half year These changes may be accomplished as follows:

28 583 trend ordinate as of January 1, 1930
0 1648 amount of increase from January 1, 1930 to July 1, 1930
28.7478 trend ordinate as of July 1, 1930

0.1648 amount of change each half year
2 number of half years in each year
0 3296 amount of change each full year

Thus we get a new trend equation based on the period 1923–1936

$$Y = 28.748 + 0.3296x$$

where x is now measured in years from July 1, 1930. Annual trend values can easily be found by following the procedure outlined above.

Monthly trend values are obtained from a monthly equation of trend. In the case of the annual equation of trend the increment rate (b) specifies the change in production which takes place in one year. To obtain the change which takes place from month to month, one-twelfth of the annual increment must be taken; where the annual increment is 0.3077 million pairs, the monthly increment is 0.3077/12, or 0.0256 million pairs. Using this increment we obtain an equation of trend

$$Y = 28.477 + 0.0256x$$

where x represents any number of months from July 1, 1929. Since ordinates of trend should generally refer to the center of the

TABLE 117

DERIVATION OF JULY ORDINATES FOR EACH YEAR—BASED ON DATA OF TABLE 111*

Month	1926	1927	1928	1929	1930	1931	1932
January							
February							
March							
April							
May							
June							
July		27 86	28 17	28.48	28.78	29.09	
August							
September							
October							
November							
December							

* Unit: million pairs.

period, in this case to the middle of the month, this equation should be so adjusted that the ordinates will refer to the fifteenth of the month rather than the first.

The ordinate for July, 1929, should be computed as of July 15, 1929, and is obtained by adding one-half the monthly increment, or 0.0128 million pairs, to the central ordinate, 28.477, to yield 28.4898 million pairs. The equation then becomes

$$Y = 28.4898 + 0.0256x$$

where x represents the number of months from July 15, 1929. When $x = 1$, the trend value for August 15, 1929, can easily be obtained from the equation; when $x = 2$, that for September 15, 1929, etc. (Table 118). As in the case of annual ordinates of

TABLE 118

DERIVATION OF MONTHLY ORDINATES—BASED ON DATA OF TABLE 111*

Month	1927	1928	1929	1930	1931
January		28 04	28.34	28.64	
February		28 06	28 36	28 67	
March		28 08	28 39	28 70	
April		28 11	28.41	28 72	
May		28.13	28.44	28 75	
June		28.16	28.46	28 77	
July		28 18	28 49	28 80	
August		28 21	28.52	28 82	
September		28 23	28 54	28 85	
October		28 26	28 57	28 87	
November		28 28	28 59	28 90	
December		28 31	28 62	28 92	
Monthly average		28 17	28.48	28 78	

* Unit: million pairs.

TABLE 119

WORK SHEET FOR FINDING ORDINATES, AND RELATIVES TO TREND*

Year and month	Actual item (1)	Ordinate of trend (2)	Relative actual-to-ordinate (3)
1929			
.....	.....	.....	...
October	37 19	28 57	130
November	27 72	28 59	97
December	22 47	28 62	78
1930			
January	26 53	28 64	93
February	25 90	28 67	90
March	28 63	28 69	100
April	29 00	28 72	101
May	24 51	28 74	85
June	23 90	28.77	83
July	24 12	28 80	86
August	28 43	28 82	99
September	29 33	28 85	102
October	27 73	28 87	96
November	18 54	28 90	64
December	17 54	28.92	61
1931			
January	19 89	28 95	69
February	23.97	28 98	82
.....	.....	.....	...
.....	.....	.....	...

* Unit: million pairs for columns 1 and 2; per cent for column 3.

trend, these calculations can be performed most efficiently on a calculating machine.

The differences between corresponding months in successive years should equal the annual increment. In addition, the monthly average trend value for each year (Table 118) will check with the annual ordinates of trend (Table 117) unless there is an error in the calculation of the monthly ordinates. A further supplementary check on the accuracy of the computations consists in dividing the difference between the Julys at the beginning (1922) and end (1936) of the interval by the number of years (15); the result should be the annual increment.

THE ELIMINATION OF TREND

The elimination of trend consists in dividing each actual item of Table 111 by the corresponding ordinate of trend of Table 118,

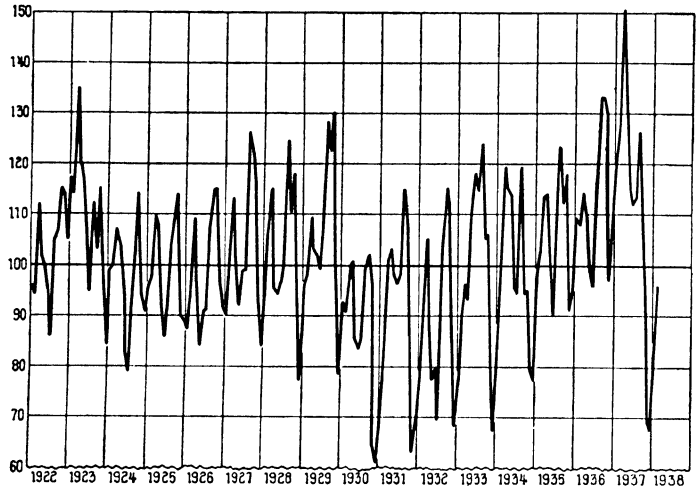


CHART 109.—Per cent ratio of actual item to ordinate of trend, for production of boots, shoes, and slippers in the United States, monthly.
(Data in column 3, Table 119.)

and expressing the result as a percentage. In actual practice this operation is facilitated if the work sheets are planned with actual items, ordinates of trend, and per cent relatives in successive columns (Table 119). These results, the *per cent relatives actual to trend* exhibit all the fluctuations of the original series except that portion due to the estimated trend. Chart 109 shows the corresponding curve; the line of trend of Chart 104 has become the

horizontal 100 per cent line, and the corresponding long-time growth of the curve of Chart 104 is absent from the curve of Chart 109.

The curve, however, continues to show the existence of seasonal variation; and the original cyclical fluctuations appear substantially unchanged.

NEGATIVE, BROKEN, AND CURVILINEAR TRENDS

Thus far the discussion has rested upon the assumption that the trend is an upward line. A downward rectilinear trend is occasionally found (Chart 110), and the analysis follows the same

CHART 110—Monthly average tax-paid withdrawals of large cigars, annually. (Data in Table *W*, page 382.)

plan except that b is a negative number. In some cases the trend line is horizontal; $b = 0$, and a is represented by the average of the actual items over the interval selected. In a long series the trend may be broken and consist of two straight lines (Chart 111), but, as economic changes do not ordinarily take place abruptly, such breaks in trend are quite exceptional. There are economic factors in which the trend appears to be curvilinear, whether concave upward or concave downward; and there may be cases in which the trend is concave upward in part of the time interval and concave downward in another part. In any particular instance, the decision as to the nature of the trend is based upon a study of the chart of actual data. The conclusion must conform to those generally accepted notions, as to the long-time tendency to growth

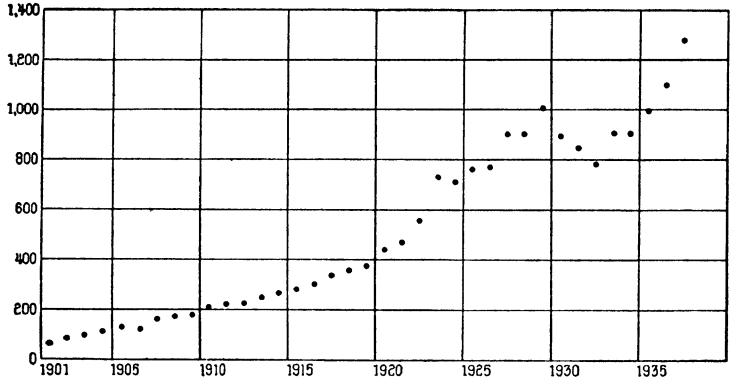


CHART 111.—Crude petroleum produced in the United States, annually.
(Unit: million barrels. Data in Table *X*, page 383.)

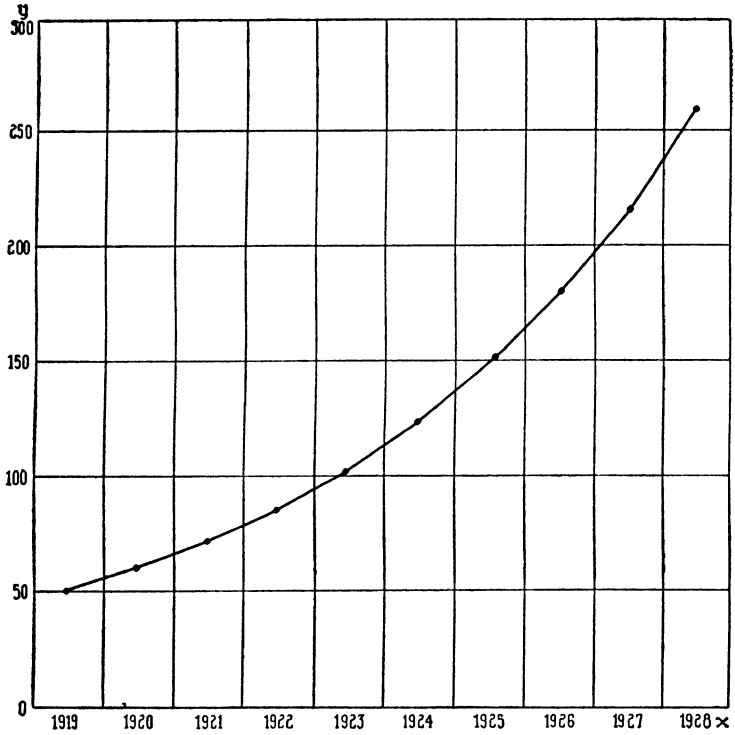


CHART 112.—Compound-interest curve, natural scale.
(Rate = 20 per cent per year.)

or decline, which are well supported by economic and historical considerations some or all of which may not be statistical.

Among the possible curvilinear trends that are most frequently needed in practice is the *compound-interest* curve (Chart 112). Whereas in the straight-line trend the growth per unit of time is a fixed *amount*, for the compound-interest curve it is a fixed *ratio*. In the one case the *absolute* change is fixed, in the other the *relative* change is fixed. Consequently, the compound-interest curve

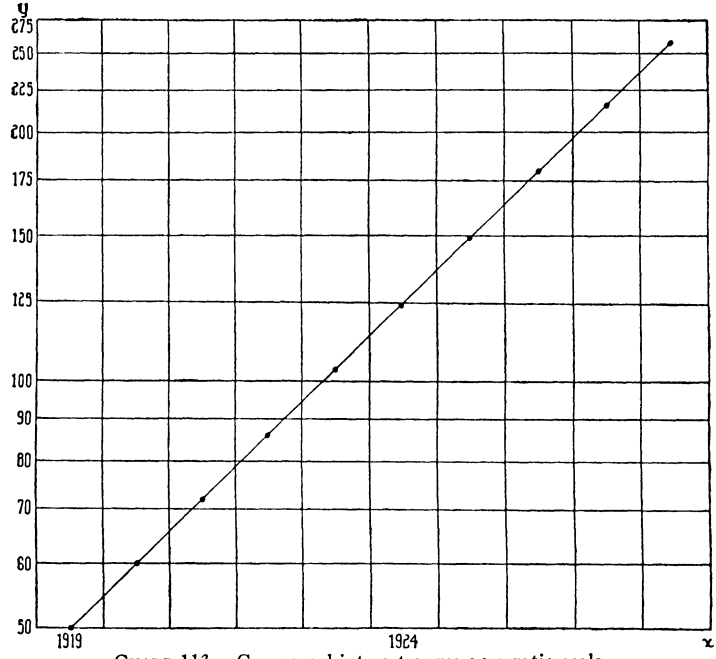


CHART 113.—Compound-interest curve on a ratio scale.

appears as a straight line on a ratio chart (Chart 113). Therefore, if a series is plotted on a ratio scale and if the trend then appears to be a straight line, the real trend of the actual data is a compound-interest curve.

Table 120 and Chart 114 show a series with this law of growth. The technique of measuring the trend, finding the ordinates, and eliminating the trend is only slightly different from that used for the rectilinear trend. The equation of the compound interest curve is

$$Y = ab^x$$

and in terms of logarithms this equation is

$$\log Y = \log a + x \log b$$

For example, suppose Y represents the monthly average output of electric power sold locally by the Hartford Electric Light Company expressed in units of million kilowatt hours, and x represents the number of years measured from July 1, 1924 as a base. A straight-line trend is actually determined for the logarithms of the

TABLE 120

OUTPUT OF ELECTRIC POWER SOLD LOCALLY BY THE HARTFORD ELECTRIC LIGHT COMPANY, HARTFORD, CONNECTICUT: ANNUAL MONTHLY AVERAGE DATA. COMPUTATION OF ORDINATES AND ELIMINATION OF TREND*

Year	Actual data	Logarithms of ordinates of trend	Ordinates of trend	Actual data with trend eliminated (%)
	(1)	(2)	(3)	(4)
1913	3 35	0 58344	3 83	87
1914	3.47	0 61839	4 15	84
1915	4 28	0 65334	4 50	95
1916	5 19	0 68829	4.88	106
1917	5 89	0 72324	5.29	111
1918	6.60	0 75819	5 73	115
1919	6.42	0 79314	6.21	103
1920	6 94	0 82809	6 73	103
1921	6 57	0 86304	7 30	90
1922	7.60	0 89799	7.91	96
1923	8.50	0 93294	8 57	99
1924	8 92	0 96789	9 29	96
1925	10 59	1 00284	10 07	105
1926	11 40	1 03779	10 91	105
1927	12 10	1 07274	11.82	102
1928	14 00	1 10769	12 81	109
1929	18 32	1 14264	13 89	132
1930	17 82	1 17759	15 05	118
1931	17 36	1 21254	16 31	106
1932	16 22	1 24749	17.68	92
1933	16 48	1 28244	19.16	86
1934	17 71	1 31739	20 77	85
1935	19.65	1 35234	22.51	87
1936	20.91	1 38729	24.39	86
1937	23.10	1.42224	26.44	87

* Unit, columns 1 and 3: million kilowatt hours. Source: *Annual Reports* of The Hartford Electric Light Company, which has granted permission for the use of the data. Monthly average output computed by the authors from the annual data obtained in the *Annual Reports*.
To obtain monthly ordinates of trend the obvious extension of the method of Table 118 would yield monthly items for column 2. The rest of the work proceeds as above.

annual monthly averages (Table 121). The equation of trend may be expressed in terms of logarithms

$$log\ Y = 0.96789 + 0.03495x$$

or in terms of actual units

$$Y = (9.287)(1.0838)^x$$

This equation shows that the rate of increase is 8.38 per cent annually.

The determination of the annual ordinates of trend in this case is most readily accomplished by use of the first equation. By use

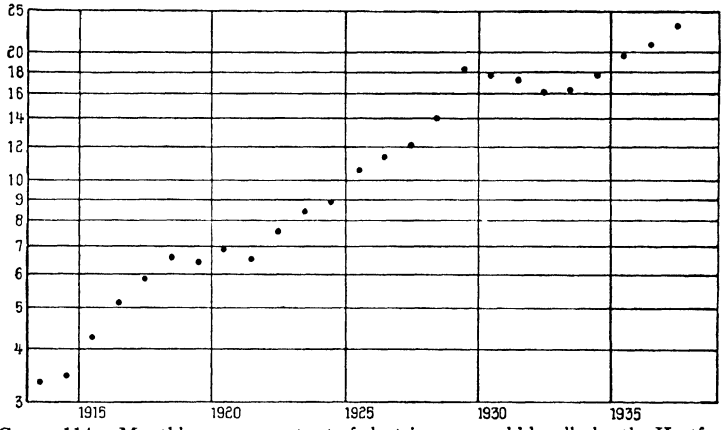


CHART 114 —Monthly average output of electric power sold locally by the Hartford Electric Light Company, annually.
(Unit, million kilowatt hours. Data in column 1, Table 120.)

of this equation the logarithms of the actual ordinates of trend can be determined by the same method used previously in calculating ordinates from the equation of straight-line trend on an arithmetic scale (column 2 of Table 120). These ordinates are the ordinates of the straight line which fits the curve on the ratio chart, and they are the *logarithms* of the ordinates of the real curved trend as it would appear on the arithmetic scale. Therefore, for each ordinate of column 2, a number (the *antilog*) is found from a logarithm table such that the given item of column 2 is the logarithm of that number. The results, entered in column 3, are the true ordinates of the compound-interest curve, which is the trend on the arithmetic-scale chart. The elimination of trend then consists in dividing each item of column 1 by the corresponding

item of column 3. The results, in percentage form, appear in column 4, and are the relatives to trend for this series.

The straight-line trend and the compound-interest trend are the two types most frequently used in elementary statistical

TABLE 121

FITTING A COMPOUND-INTEREST CURVE TO THE OUTPUT OF ELECTRIC POWER SOLD LOCALLY BY THE HARTFORD ELECTRIC LIGHT COMPANY

Year	Y Monthly average output (million kilowatt hours)	log Y	x	$x \log Y$	
				−	+
1913	3 35	0 52504	−11	5 77544	
1914	3 47	0 54033	−10	5 40330	
1915	4 28	0 63144	− 9	5 68296	
1916	5 19	0 71517	− 8	5 72136	
1917	5.89	0 77012	− 7	5.39084	
1918	6.60	0 81954	− 6	4 91724	
1919	6 42	0 80754	− 5	4 03770	
1920	6 94	0 84136	− 4	3 36544	
1921	6 57	0 81757	− 3	2 45271	
1922	7 60	0 88081	− 2	1 76162	
1923	8 50	0 92942	− 1	0 92942	
1924	8 92	0 95036	0		
1925	10 59	1 02490	+ 1		1 02490
1926	11 40	1 05690	+ 2		2 11380
1927	12 10	1 08279	+ 3		3.24837
1928	14 00	1 14613	+ 4		4.58452
1929	18 32	1 26293	+ 5		6 31465
1930	17 82	1 25091	+ 6		7 50546
1931	17 36	1 23955	+ 7		8 67685
1932	16 22	1 21005	+ 8		9 68040
1933	16 48	1 21696	+ 9		10 95264
1934	17 71	1 24822	+10		12 48220
1935	19 65	1 29336	+11		14 22696
Total		22 26140		−45 43803	+80 81075

Net total = +35 37272

$$a = \frac{22.26140}{23} = 0.96789. \quad b = \frac{35\ 37272}{1012} = 0.03495$$

work, although one occasionally encounters an economic series in which neither of the above trends will fit the chart of actual items. In such cases another type of curve must be selected to measure trend. One such curve is the parabola. If, when a series is plotted on a ratio scale, it appears to have a trend which is concave

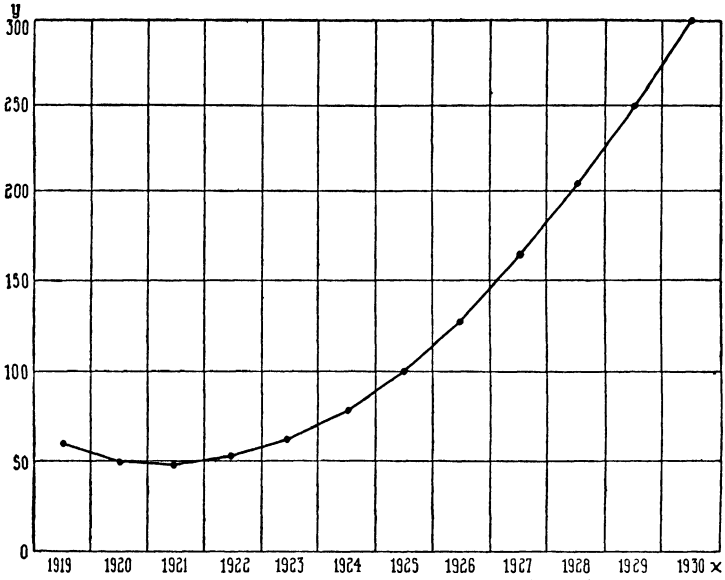


CHART 115.—Concave upward parabola, with minimum (vertex) in 1921 on a natural scale.

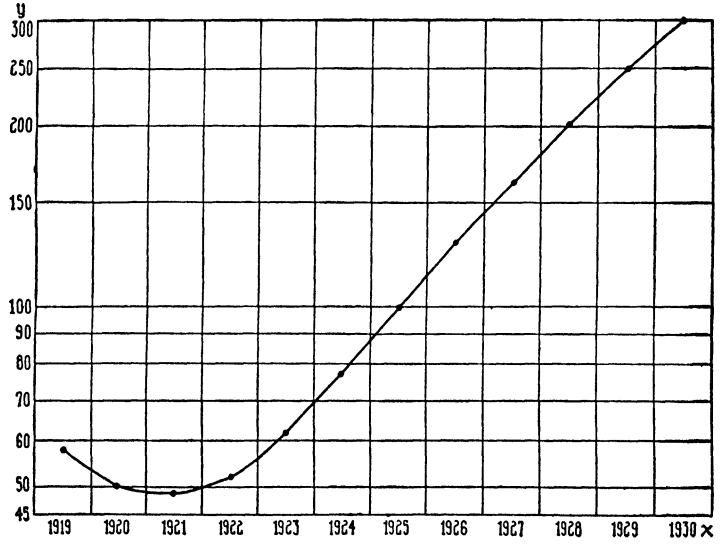


CHART 116.—Concave upward parabola, with minimum (vertex) in 1921 on a ratio scale.

upward, the compound-interest law of growth will not fit the series; in many such cases a parabola will best fit the series (Chart 115, arithmetic; Chart 116, logarithmic). For the parabola the upward concavity strictly holds only for the flatter part of the curve as plotted on the ratio scale. The fitting of the parabola by computation involves more elaborate calculations than those needed for the straight line and is not shown here. Moreover, the labor in deriving the ordinates is also increased.

If, when a series is plotted on an arithmetic scale, the trend appears to be concave downward, the law of growth is likely to be

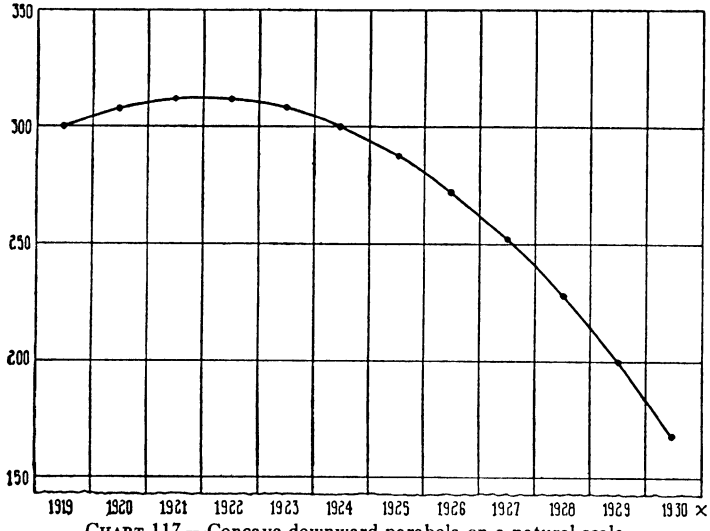


CHART 117.—Concave downward parabola on a natural scale.

a parabola with curvature in the direction opposite to that discussed above (Chart 117). The calculations in this case follow exactly the same lines as for the parabola which is concave upward.

The decision as to the *form of the trend* constitutes a further discretionary act on the part of the investigator, the decision on the *length of the interval* being the first. In practice these decisions cannot be made independently—the one is linked with the other. Both are reached by a study of the chart of actual items. Generally, in making this chart, a ratio scale should be used. If then the trend appears straight, the real trend is of the compound-interest type; if it appears concave upward, the real trend is likely to be a parabola concave upward; if it appears moderately concave down-

ward, the real trend is likely to be a straight line; and, if it appears strongly concave downward, the real trend is likely to be a parabola. In the last two cases, if doubt exists, the curve should be plotted (at least in the form of annual figures) on the arithmetic scale. Although other types of mathematical equations may occasionally be needed to represent the trend in a specific series, such instances are rare; most cases can in fact be handled by the straight line or the compound-interest curve.

CHAPTER XXI

SEASONAL VARIATION

GRAPHIC EVIDENCE OF SEASONAL TENDENCY

The second measurable component of the total fluctuation is the seasonal variation. It should be kept in mind that one of the primary objects of the analysis of a particular economic series is to isolate, for separate study, the cyclical fluctuation; and, in order to do this, other components of the total movement must be measured and eliminated. The preceding chapter dealt with the measurement and elimination of secular trend, and the present chapter treats of the measurement and elimination of the seasonal variation.

Here also the first step consists in an examination of the chart of actual data for evidence of the existence of seasonal variation. If the chart shows on the whole a persistent tendency for the items to be higher in certain months of each year and lower in certain other months, the assumption is warranted that there is a seasonal movement, and an arithmetic process such as that described below is used to estimate its precise form and extent.

For this purpose a chart on the ratio scale is ordinarily desirable. Chart 118 exhibits the data of Table 122, on daily average production of Portland cement. Here the upward movement in the early months of the year and the high level of production from June to September, followed by a sharp drop in the later months, are readily apparent. The usual seasonal swing is, however, altered in form at times by the cyclical sweep, for example, in the last half of 1931 and in the years 1933–1935. In this series there is no doubt of the existence of a seasonal swing, and a rough estimate of its form and extent can be obtained from the chart.

The fact that, on the ratio scale, the seasonal movement has substantially the same extent in years when production is at a relatively high level as when it is at a low level (during the period 1932–1937) suggests that the seasonal variation is essentially a *relative* phenomenon. In other words, the same seasonal influence tends to produce the same relative change in the given economic factor, and produces therefore a larger absolute deviation at higher

TABLE 122

DAILY AVERAGE PRODUCTION OF PORTLAND CEMENT*

Month	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938
January	285 7	254.5	266 4	315 1	318 7	274 1	212.7	162 2	95 4	121 9	103 2	117.7	213 4	146.2
February	294 8	276 1	263 5	303 4	304 4	291 5	211 4	136 9	98 2	148 8	109 0	119 8	208 5	139.9
March	356 0	335.2	369 4	329 8	321.6	362.1	266 0	156 4	118 8	169 5	138 6	171 3	272 4	
April	460 2	414.7	468.3	448 9	458 3	450 7	374 8	182 6	139 4	218 1	204.5	287 1	346 7	
May	500 2	532.7	538 8	558.3	521 0	556 2	451 9	223 0	202 0	275 6	265 2	358.2	375.3	
June	512.9	562.2	574.1	583.2	560 1	574 6	470 6	264 0	260.1	293 8	290.4	379.2	372.1	
July	504 6	552.8	561.5	563.7	558 5	550 8	448 4	247.1	277 7	262.7	258.7	371.1	374.1	
August	529 6	548 3	590.8	605 1	599 5	575.0	437 1	252 8	265 2	252 9	233 4	405 5	383.7	
September	531 3	552 4	583 5	596 1	574 1	537 5	403.0	273 6	187 9	256.0	239 1	411 6	374.1	
October	516 0	535 4	554 1	565 6	539 7	464 9	347.2	256.1	162 5	215 3	242.3	402.3	366.9	
November	455 2	473 1	481 6	502 3	468 4	369 9	272 0	215 4	155 7	192 6	236.4	365 9	308.3	
December	345.7	347.1	387 1	393 2	361.8	273 6	192 7	137 0	113 7	143 4	187.2	289 4	227.3	

* Unit: thousand barrels. Daily average production has been computed from monthly totals by dividing each monthly total by the number of calendar days in each month. Source of monthly totals: "Mineral Resources of the United States," 1926–1930, Part II, "Statistical Appendix to Minerals Yearbook," 1932–1933–1935, "Minerals Yearbook," 1936–1937, and "Monthly Cement Statement," 1937–1938, Washington, U. S. Department of Commerce and U. S. Department of the Interior.

than at lower levels. On the ratio scale these equal relative deviations appear equal. This uniformity of relative change is a general property of seasonal variation, although there are excep-

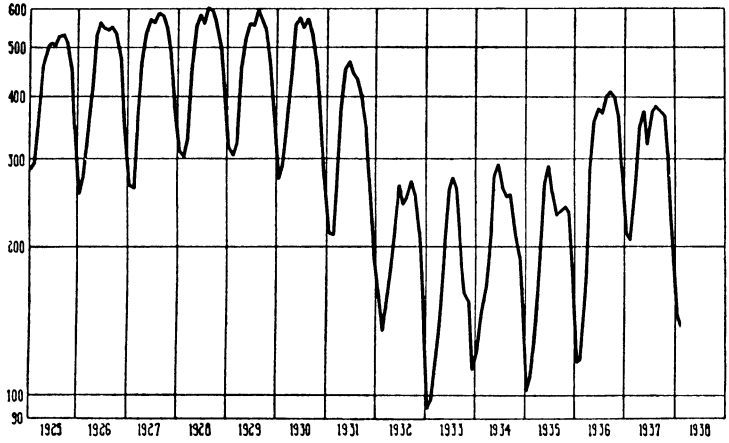


CHART 118.—Daily average production of Portland cement in the United States, monthly.
(Unit: thousand barrels. Data in Table 122. Ratio scale.)

tional series in which it does not hold; it is for this reason that the actual items should be plotted on a ratio scale to yield a visual test of seasonal variation. In practice, a single chart on the ratio

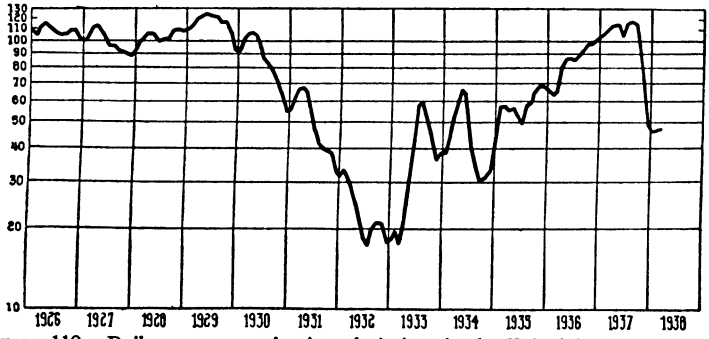


CHART 119.—Daily average production of pig iron in the United States, monthly.
(Unit: thousand long tons. Data in Table 139, page 354. Ratio scale.)

scale is generally adequate to determine both the nature of the trend (see page 324) and the existence of seasonal variation.

In many instances a clear conclusion as to the seasonal variation cannot be reached by inspection of the curve (Chart 119).

For these cases a more elaborate graphic comparison may be made by superposition of all the subsequent annual sections of the curve upon the section for the first year, either by plotting each year on

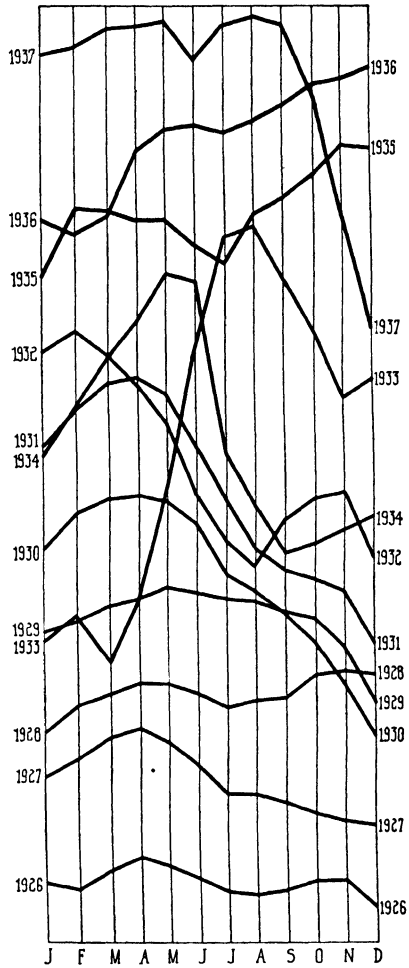


CHART 120.—Annual segments of the curve of Chart 119, arranged on a single 12 months as base.

a transparent sheet and superposing the sheets or by plotting all the annual curves on a single sheet with a common base interval of one year, as in Chart 120. This device may help to bring out such

similarities of seasonal change as exist. If, however, doubt still remains as to the existence of any seasonal movement in the series, the arithmetic process presented below furnishes the ultimate test. It is well to note, finally, that there are some series in which the chart clearly shows that no seasonal changes whatever take place. In such cases, of course, the "seasonal index" is 100 per cent for each month in the year.

THE DISTRIBUTION OF LINK RELATIVES

There are many arithmetic processes for determining the estimate of seasonal variation. The most satisfactory general process is the *median-link-relative* method developed by Professor Persons.[1] The various steps in this method will be shown by working out in detail the calculations for the series of Table 122 (data for the January, 1925, to December, 1937, interval). As in the case of the calculation of secular trend, the interval chosen should include more than one cycle. A period of ten years is ordinarily sufficient to obtain a significant measure of seasonal variation, unless the form of the seasonal movement is undergoing change. In no case, however, should the period selected be less than seven years.

The first operation consists in computing, from the actual data, the link relative (in precisely the sense of Chap. XVIII) for each month from January, 1925, to December, 1937. The link relative for January, 1925/December, 1924 (daily average production of Portland cement in December, 1924, was 336.7 thousand barrels) is 285.7/336.7 = 85 per cent; that for February, 1925/January, 1925 is 294.8/285.7 = 103 per cent; that for December, 1937/November, 1937 is 227.3/308.3 = 74 per cent. The results are exhibited in Table 123. The calculations in this instance are carried only to the nearest per cent. In some cases the relatives should be carried to one decimal place, but the decision to do this must rest partly upon the precision of the original data and partly upon the nicety with which it is expected to determine the seasonal indexes.

These relatives show the manner in which daily average production of Portland cement varies from month to month throughout the interval under examination. They show, for example, how each of the thirteen Januarys was related to the preceding December. In order to find the seasonal movement it is necessary to ascertain how, *on the average*, January is related to December

[1] *Review of Economic Statistics*, Prel. Vol. I, January, 1919, pp. 18–31.

throughout the thirteen years and, similarly, how, on the average, each of the other eleven months is related to the preceding month of the year. It would be possible to compute the average ratio of January to December directly from Table 123, according to any one of the definitions of *average* developed in previous chapters, but such a computation would not indicate definitely whether the result has any true significance. The question, whether or not the resulting average is truly typical, would remain.

TABLE 123

LINK RELATIVES FOR DAILY AVERAGE PRODUCTION OF CEMENT*

Year	Jan. / Dec.	Feb. / Jan.	Mar / Feb.	Apr. / Mar.	May / Apr.	June / May	July / June	Aug. / July	Sept. / Aug.	Oct. / Sept.	Nov / Oct	Dec. / Nov.
1925	85	103	121	129	109	102	98	105	100	97	88	76
1926	74	108	121	124	128	106	98	99	101	97	88	73
1927	77	99	140	127	115	107	98	105	99	95	87	80
1928	81	96	109	136	124	104	97	107	99	95	89	78
1929	81	96	106	142	114	107	100	107	96	94	87	77
1930	76	106	124	124	123	103	96	104	93	86	80	74
1931	78	99	126	141	121	104	95	97	92	86	78	71
1932	84	84	114	117	122	118	94	102	108	94	84	64
1933	70	103	121	117	145	129	107	96	71	87	96	73
1934	107	122	114	129	126	107	89	96	101	84	89	74
1935	72	106	127	148	130	110	89	90	102	101	98	79
1936	63	102	143	168	125	106	98	109	101	98	91	79
1937	74	98	131	127	108	99	101	103	97	98	84	74

* Unit: one per cent. Based upon data of Table 122.

The important point is that the thirteen relatives, January–December, constitute a frequency series, and that the appropriateness of an average for this series depends upon the form of the distribution: questions of dispersion and skewness at once arise. To exhibit the essential features of the twelve distributions, one for each month, a frequency series of the thirteen relatives should be made for each month. Moreover, to facilitate comparisons between successive months, the twelve frequency series should be arranged in adjacent columns with a single scale of class intervals. The class interval is usually taken as 1 per cent, and the centers of the intervals are taken as integers. But there are cases—as in exchange rates—in which the seasonal swing is ordinarily slight, and for which a class interval of 0.1 per cent is needed. The range of the per cent scale is determined by the high and low relatives of Table 123, with the understanding that a small number of the relatives for any one month may fall in the "all over" or "all

TABLE 124

MULTIPLE-FREQUENCY DISTRIBUTION OF LINK RELATIVES FOR DAILY AVERAGE PRODUCTION OF PORTLAND CEMENT, 1925 TO 1937*

Relatives	Jan / Dec	Feb / Jan	Mar / Feb	Apr. / Mar	May / Apr	June / May	July / June	Aug / July	Sept / Aug	Oct / Sept	Nov / Oct	Dec / Nov
Medians	77 0	102 0	121 0	129 0	123 0	106 0	98 0	103 0	99 0	95 0	88 0	74 0
Over 140												
140												
139												
138												
137												
136												
135												
134												
133												
132												
131												
130												
129												
128												
127												
126												
125												
124												
123												
122												
121												
120												
119												
118												
117												
116												
115												
114												
113												
112												
111												
110												
109												
108												
107												
106												
105												
104												
103												
102												
101												
100												
99												
98												
97												
96												
95												
94												
93												
92												
91												
90												
89												
88												
87												
86												
85												
84												
83												
82												
81												
80												
79												
78												
77												
76												
75												
74												
73												
72												
71												
70												
69												
68												
67												
66												
65												
64												
63												
62												
61												
60												
Under 60												

* Data in Table 123.

under" group. The twelve frequency series are arranged in this way and form the *multiple-frequency table* (Table 124).

This table shows how, in each column, the items for a particular month are distributed, and therefore indicates how typical a particular average is for the series of that column. The closeness of cluster of the scores within each column is of chief importance in this connection.

A second important feature of the distribution is the *deviation from normal* for each of the twelve months. Inspection of the table reveals whether the group of scores for a particular month tends on the whole to lie above or below 100 per cent, and roughly how great the deviation from 100 per cent is. The importance of this consideration is apparent: as these are *link* relatives, the tendency for the entire group of links for a particular month (say March) to lie above 100 per cent must reflect the fact that such a month is seasonally above the preceding month (for example, March is seasonally above February). On similar grounds, if the link-relative group for each of several successive months deviates moderately in a given direction from 100 per cent, the cumulative effect is likely to be as significant as the effect of a large deviation for a single month.

A third aspect of Table 124 concerns the *displacement* between adjacent columns. Because the scores represent link relatives, the displacement—comparative difference in level among the groups, each considered as a whole—indicates roughly the rate of change of the seasonal movement. For example, if the displacement were zero, as it would be if the two adjacent groups of relatives were on the same level (September and October of Table 125), the typical link for the second month would be equal to that for the first: the ratio of increase for the second above the first would be the same as for the first above the preceding month (October/September would equal September/August). On the other hand, if the second group is clearly displaced from the first (as April is displaced above March in Table 124), the typical link relative is increasing from the first to the second month: the second month is a larger percentage of the first than the first is of the preceding.

The direction and amount of the displacement between pairs of adjacent months thus indicate the changes in the direction or rate of the seasonal shift. Thus Table 125 shows the hypothetical scores to represent the seasonal movement of Chart 121. For the first three months, in which the seasonal movement of Chart 121 is upward at an approximately uniform rate, there is no displace-

ment; then, as the seasonal movement flattens off during the next two months, there is a moderate downward displacement; next, as the seasonal swing reverses its direction, there is a single large downward displacement; and, as the seasonal movement straightens out again into the decline at uniform rate during September and October, the displacement reduces to zero again. The changes for this illustrative seasonal curve are unusually regular,

TABLE 125

MULTIPLE-FREQUENCY DISTRIBUTION OF LINK RELATIVES FOR THE HYPOTHETICAL SERIES HAVING SEASONAL MOVEMENT SHOWN IN CHART 121

Relatives	Jan. Dec.	Feb. Jan.	Mar. Feb.	Apr. Mar.	May Apr	June May	July June	Aug. July	Sept. Aug	Oct. Sept.	Nov. Oct.	Dec. Nov.
Over 120												I
120												I
119												I
118												I
117												III
116												
115												II
114												
113												I
112												I
111											I	I
110			I									
109		I										
108		I		I								
107	II	II	I								I	
106	I		II									
105	I	II	I	I	I							
104	III	II	II	II	I						II	
103	I	I	II	I							I	
102	II	II		II	I						III	
101	I		I	III	III	I					I	
100	I	I			II					I	I	
99			I	I		I	I		I			
98			I		I	I	I		I	I	I	
97				II	II	III				I	I	
96				I		I	III	I	I			
95					I	II	I	II	II	II		
94					I	I	II	I	III			
93						I	II	I	I	III		
92								II		II		
91						II	I	III	I	I		
90							I	I	I			
Under 90							I	II	I	I		

but they serve to suggest the significance of the displacement found in the multiple-frequency table.

The three properties, closeness of cluster, deviation from normal, and displacement, must be considered simultaneously in examining the multiple-frequency table. The basic purpose of this examination is to determine whether the seasonal movement exists and can be measured precisely. With almost no exception, seasonal variation exists if there is clear deviation from normal for some or all of the months. Whether such seasonal variation can be measured precisely depends upon the structure of the table as a whole. If the cluster is very close, determination is precise, even if deviations from normal are not large. If deviations from normal are large, good precision is attainable, even if the cluster is not

close. If the cluster is not close and deviations from normal are not considerable, the measurement cannot be precise. In general, the precision is lower according as displacement is more widely present. In extreme cases, where, for instance, the cluster is not close and the "typical" deviations from normal are small (Table 126), the conclusion is that no seasonal variation exists or that such as does exist cannot be measured accurately.

Instances occur in which the evidence of the table suggests that the seasonal variation exists and can be measured, but not with high precision. For such cases it is often necessary to abandon the

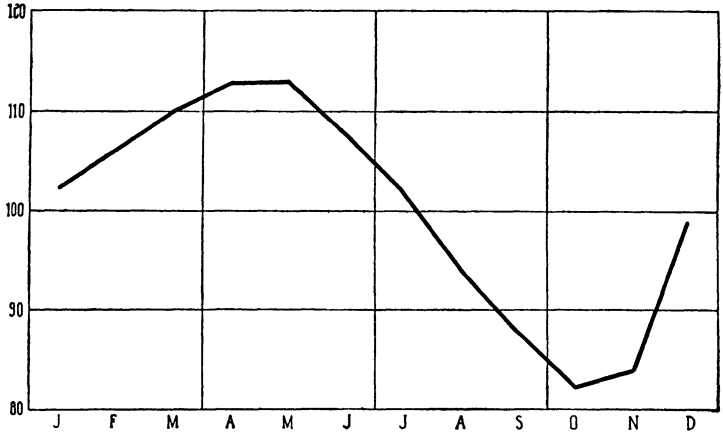


CHART 121.—Hypothetical seasonal movement for the data of Table 125.

attempt to measure the seasonal movement to the nearest per cent, and it may even be wise to construct a new multiple-frequency table, with a class interval of 2 per cent or even 5 per cent (Table 127).

THE MEDIAN LINK RELATIVES

Returning to Table 124, the decision seems warrantable that the seasonal variation exists and can be measured precisely. It is possible to secure from the multiple-frequency table some estimate of the hidden seasonal wave itself. For example, the production for March clearly must be above that for February, as the March group of links is clearly above 100 per cent. Somewhat similar deductions are possible for some of the other months. Until the investigator becomes skilled, however, this use of the table may

TABLE 126

MULTIPLE-FREQUENCY TABLE OF LINK RELATIVES OF A HYPOTHETICAL SERIES WHICH HAS NO SEASONAL MOVEMENT, OR A SEASONAL MOVEMENT SO SLIGHT AND IRREGULAR THAT IT CANNOT BE MEASURED

Relatives	Jan.–Dec.	Feb.–Jan.	Mar.–Feb.	Apr.–Mar.	May–Apr.	June–May	July–June	Aug.–July	Sept.–Aug.	Oct.–Sept.	Nov.–Oct.	Dec.–Nov.
Over 120												
120												
119												
118							I					
117												
116			I									
115							I					
114												
113							I					
112	I		I	I								
111					I							
110												
109	I			I	I		I			I	II	
108		I		I					I			
107	I		I			I					I	I
106					I	II		I	II	I		
105	II	I	I	I	I		I	I		I	I	II
104		I	I		I	I			I		I	
103	I		II		I			I		I		
102		I		I			I	I	I	I	I	
101	I	I	I		II				II		I	I
100		II				I		I			II	
99	I		I	I	I	I	I	II	I	II		
98		I	I	I								III
97	I			I		I	I	I		I	I	I
96	I	I	I		I				I		I	
95		I		I			I	I		I		
94					I	I	I		I			
93	I			I								I
92								I		I		I
91		I								I		
90				I		I				I		
Under 90						I		I				I

TABLE 127

MULTIPLE-FREQUENCY TABLE OF LINK RELATIVES FOR THE RATE OF INTEREST ON CALL LOANS IN THE NEW YORK STOCK EXCHANGE, FEBRUARY, 1903 TO JANUARY, 1914*

Relatives	Jan.–Dec.	Feb.–Jan.	Mar.–Feb.	Apr.–Mar.	May–Apr.	June–May	July–June	Aug.–July	Sept.–Aug.	Oct.–Sept.	Nov.–Oct.	Dec.–Nov.
Over 150			I	I					III	IIII	I	III
150								I		I		I
145			II				I				II	
140										I		
135						I				I		
130									III			
125		I	I	I		I			II	I		I
120											II	I
115				I					I	I		
110					II		I				I	II
105		I	III	I		II	I	I		I	I	II
100			II	II	II	II	II	III	I	I	I	I
95	II	II	I	I	II				I			
90				I	I	I	III	I				
85						I		II				
80			I	II	I	II	II	I				
75		I			I	II						
70	I	II		I								
65								II				
60	II	I			I						I	
55		I								I		
50	III	I										
Under 50	III	I		I	I							

* Based on data in Table *Y*, page 383.

prove misleading. When a clear case exists, as for Table 124, the next essential step is the computation of the seasonal index.

To get the relation which holds, *on the average*, between January and December, some summary of the thirteen individual January/December ratios is necessary. The objection to the arithmetic average (mean) is that these frequency series are precisely of the sort for which the mean is not likely to be typical. The reason is that the original data, from which the link relatives were computed, may be affected with irregular deviations; and these irregular deviations would appear in the frequency table as extremely high or extremely low link relatives. The fact that the mean is peculiarly subject to the influence of extreme items renders its use undesirable in this problem; for, if it were used, the resulting index might well be governed by the exceptional and irregular items rather than by the seasonal movement which it is sought to measure. Therefore, the type of average selected is the median, and the medians for the twelve calendar months are presented at the top of Table 124.

If the concentration of the individual link relatives about the medians, for the particular months, were not so close as it is in this case, a compromise in the use of the median would be desirable: for an odd number of monthly link relatives, taking the average of the middle three or five items (instead of the middle one); and for an even number, taking the average of the middle four or six items (instead of the middle two items). The need for this modification generally appears when there is, in certain monthly frequency columns, a partial isolation of the median; for example, when there is no frequency in the per cent intervals just below or just above the median. If this device appears necessary for one or more months of a table, it should be applied to all.

In the simple case treated above, straightforward examination of the frequency table yields the medians. The medians represent the *typical* relation of each of the twelve months to the preceding month. They do not pertain to any particular year in the specific time interval 1925–1937; they belong to a hypothetical year in which the changes from one month to the next are those which take place "usually." These median link relatives are not in themselves satisfactory indexes of seasonal movement, for they compare each month with that immediately preceding. Indexes on a single base are preferable. In order, therefore, to convert the medians to a useful form, further computations are necessary. The obvi-

ous step is to construct chain indexes from the link indexes, in accordance with the definition given in Chap. XVIII.

THE DETERMINATION OF SEASONAL INDEXES

The chain index for any month is simply the product of all the preceding link indexes of the year; so that the chain index for December will be obtained by multiplying together all twelve of the links. The actual calculation may be done by the use of logarithms, and the details are presented in Table 128. The

TABLE 128

COMPUTATION OF THE INDEX OF SEASONAL VARIATION FOR DAILY AVERAGE PRODUCTION OF PORTLAND CEMENT IN THE UNITED STATES—LOGARITHMIC METHOD

Monthly comparison	Link relative		Month	Chain index, uncorrected (December of preceding year = 100)		Correction (applied to logarithm)	Chain index, corrected		Index of seasonal variation (av. for year = 100)
	Actual	Log		Log	Actual value		Log	Actual value	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Jan./Dec.	0 77	9 8865–10	Jan.	9 8865–10	0 770	0 0004	9 8869–10	0 771	59
Feb./Jan.	1 02	0 0086	Feb	9 8951–10	0 785	0 0009	9 8960–10	0 787	61
Mar./Feb	1 21	0 0828	Mar.	9 9779–10	0 950	0 0013	9 9792–10	0 953	73
Apr./Mar	1 29	0 1106	Apr	0 0885	1 226	0 0018	0 0903	1 231	95
May/Apr.	1 23	0 0899	May	0 1784	1 508	0 0022	0 1806	1 516	117
June/May	1 06	0 0253	June	0 2037	1 598	0 0026	0 2063	1 608	124
July/June	0 98	9 9912–10	July	0 1949	1 566	0 0031	0 1980	1 578	122
Aug./July	1 03	0 0128	Aug	0 2077	1 613	0 0035	0 2112	1 626	125
Sept./Aug	0 99	9 9956–10	Sept	0 2033	1 597	0 0040	0 2073	1 612	124
Oct./Sept.	0 95	9 9777–10	Oct	0 1810	1 517	0 0044	0 1854	1 532	118
Nov./Oct	0 88	9 9445–10	Nov.	0 1255	1 335	0 0048	0 1303	1 350	104
Dec./Nov.	0 74	9 8692–10	Dec.	9 9947–10	0 988	0 0053	0.0000	1 000	77
		9 9947–10		or −0 0053				15 564 av. = 1 297	1,199

calculation may also be done arithmetically, as shown in Table 129. In column 2 of Table 128 are given the medians, in column 3 their respective logarithms. For this purpose, it is well to treat the medians as decimal ratios rather than as percentages. In column 5 appear the logarithms of the chain indexes for each month and in column 6 are given the actual values of these chain indexes. Obviously the total of column 3 is the logarithm of the chain index for December in column 5.

The chain indexes are on the preceding December as base, since the first link (that for January) was based on the preceding December. If the seasonal effect were the only cause of variation, and if the median link relatives were perfectly accurate measures of the seasonal movement, the chain index for December based on

the preceding December would obviously be 100 per cent. Its logarithm should, therefore, be 0.0000. The fact that it is not 0.0000, but 9.9947 − 10 (or −0.0053), follows partly from the confusing influence of other types of fluctuation (chiefly secular trend) and partly from slight inaccuracies unavoidable in a method which cannot be more than approximate.

The chain index for December, which has 9.9947 − 10 as its logarithm, is 98.8 per cent; and, in order to alter the chain index for each month so that it will represent seasonal variation and nothing else, it is necessary to distribute this discrepancy of −1.2 per cent among the twelve monthly chain indexes in such a way that the logarithm of the corrected chain index for December will be zero. The process of distribution is an indirect one. In fact, one-twelfth of 0.0053, or 0.0004, is added to the logarithm of the uncorrected chain index for January; two-twelfths of 0.0053, or 0.0009, is added to the logarithm of the uncorrected chain index for February, etc.; the corrections appear in column 7, and the corrected logarithms of column 8 are the logarithms of those values which we estimate the chain indexes for each month would have had if they had reflected seasonal variations alone. An obvious modification follows if the discrepancy is positive instead of negative: if the logarithm of the December uncorrected chain index had been 0.0048 instead of −0.0053, the items of column 8 would have been obtained by *subtracting* one-twelfth 0.0048 from the January uncorrected chain index in column 5, two-twelfths of 0.0048 from the February uncorrected chain index, etc.

In column 9 the chain indexes which represent seasonal variation alone are all on a single base, the December of the preceding year. It is a distinct advantage to have the average for the current year as the base. A final adjustment of the indexes consists in shifting the base from December of the previous year to the average for the current year: each item of column 9 is divided by the average (one-twelfth of 15.564, or 1.297) of the twelve items, and the results are the *adjusted indexes of seasonal variation.* They are entered and rounded off to the nearest per cent, in column 10.

The total of the monthly indexes of column 10 theoretically should equal 1,200. Because of the fact that the indexes are usually expressed as percentages and are rounded off to the nearest per cent, the total may be slightly greater or less than this figure; here the total is 1,199. The total of the twelve seasonal indexes, however, should not differ from 1,200 by more than 1 or 2 in either

direction. In some cases when the total is less than 1,200, the individual indexes are rounded off to make their total equal this amount. Such rounding off is done in the calculation of the seasonal indexes for monthly average production of boots and shoes in Table 137.

The arithmetic method for the calculation of the seasonal index is illustrated in Table 129. In column 2 are presented the median link relatives, and the chain indexes for each month, with December of the preceding year as base, are given in column 4. As was

TABLE 129

COMPUTATION OF THE INDEX OF SEASONAL VARIATION FOR DAILY AVERAGE PRODUCTION OF PORTLAND CEMENT—ARITHMETIC METHOD

Monthly comparison	Typical link relative (%)	Month	Chain index uncorrected (Dec. of preceding year = 100) (%)	Correction	Chain index corrected (Dec. of preceding year = 100) (%)	Index of seasonal variation (av. for year = 100) (%)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Jan./Dec.	77.0	Jan.	77 0	0 1	77.1	60
Feb /Jan.	102 0	Feb.	78 5	0 2	78 7	61
Mar /Feb.	121 0	Mar.	95 0	0 3	95 3	74
Apr /Mar	129 0	Apr.	122 6	0 4	123 0	95
May/Apr.	123 0	May	150 8	0 5	151 3	117
June/May	106 0	June	159 8	0 6	160 4	124
July/June	98 0	July	156 6	0 7	157 3	122
Aug /July	103 0	Aug.	161 3	0 8	162 1	125
Sept /Aug.	99 0	Sept.	159 7	0 9	160 6	124
Oct /Sept.	95 0	Oct	151 7	1 0	152 7	118
Nov./Oct.	88 0	Nov.	133 5	1 1	134 6	104
Dec./Nov.	74 0	Dec.	98 8	1 2	100 0	77
					1,553.1*	1,201
					av. = 129 4	

pointed out in the discussion of the logarithmic method above, the December chain index should equal 100 per cent if seasonal variation alone is responsible for the variations in the monthly chain indexes. The negative discrepancy of 1.2 per cent is distributed arithmetically among the twelve monthly chain indexes, one-twelfth of 1.2, or 0.1, to January, two-twelfths of 1.2, or 0.2, to February, etc., and these corrections appear in column 5. In column 6 are given the corrected chain indexes with December as base, and the final seasonal indexes for each month with the average month of the year as a base are given in column 7.

TABLE 130

DAILY AVERAGE PRODUCTION OF PORTLAND CEMENT IN THE UNITED STATES ADJUSTED FOR SEASONAL VARIATION, MONTHLY*

	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938
January	484 2	431 4	451 5	534 1	540 2	464 6	360 5	274 9	161.7	206 6	174 9	199 5	361 7	247 8
February	483 3	452.6	432 0	497.4	499 0	477 9	346 6	224 4	161 0	243 9	178 7	196 4	341 8	229.3
March	487 7	459 2	506 0	451 8	440 5	496 0	364 4	214 2	162 7	232 2	189.9	234 7	373 2	
April	484 4	436 5	492 9	472 5	482 4	474 4	394 5	192 2	146 7	229 6	215.3	302 2	364 9	
May	427 5	455 3	460 5	477 2	445 3	475 4	386 2	190 6	172 6	235.6	226 7	306 2	320 8	
June	413.6	453 4	463 0	470.3	451.7	463 4	379 5	212 9	209.8	236.9	234 2	305.8	300.1	
July	413 6	453 1	460 2	462 0	457 8	451 5	367 5	202.5	227 6	215.3	212 0	304.2	306 6	
August	423 7	438 6	472 6	484 1	479 6	460 0	349.7	202 2	212.2	202.3	186 7	324.4	307.0	
September	428 5	445 5	470 6	480 7	463 0	433 5	325 0	220 6	151 5	206 5	192.8	331.9	301.7	
October	437.3	453 7	469 6	479 3	457 4	394 0	294 2	217 0	137.7	182 5	205 3	340 9	310.9	
November	437 7	454 9	463 1	483 0	450 4	355 7	261 5	207 1	149 7	185.2	227 3	351.8	296.4	
December	449 0	450 8	502 7	510.6	469 9	355 3	250 3	177.9	147 7	186 2	243.1	375.8	295 2	

* Unit: thousand barrels. Based upon data of Tables 122 and 128.

The slight differences which appear in the seasonal indexes here as compared with those in Table 128 (the indexes for January and March) are caused by the distribution of the correction factor —a logarithmic distribution in Table 128 and an arithmetic distribution in Table 129—and the rounding off of the seasonal indexes to percentage figures. The logarithmic method of calculating seasonal indexes is to be preferred, although the results obtained by the two methods in many cases are the same. In the calculations which follow the seasonal indexes obtained in Table 128 are used.

THE ELIMINATION OF SEASONAL VARIATION

The actual level of the daily average production of Portland cement in any one month is made up of secular trend, seasonal

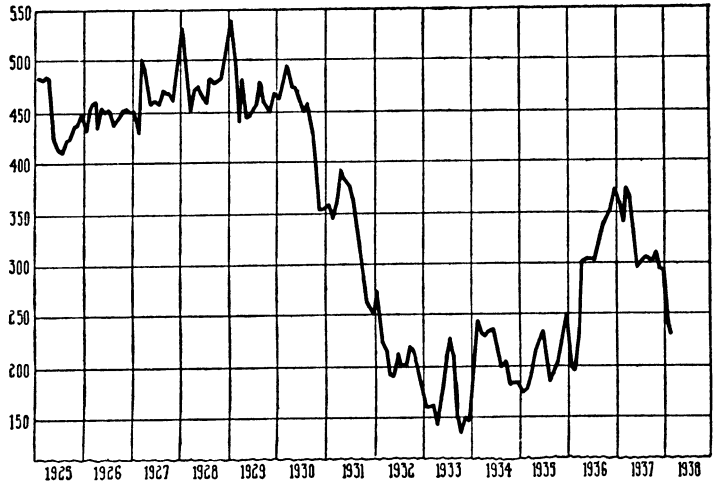


CHART 122.—Daily average production of Portland cement adjusted for seasonal variation.

(Unit: thousand barrels. Data in Table 130.)

variation, cyclical variation, and random, or irregular, influences. The elimination of seasonal variation involves the division of the daily average production of Portland cement in each month by the appropriate seasonal index for that month. Each January production figure is divided by 59 per cent in order to eliminate the seasonal factor which is usual in the production of that month. Since production in January is ordinarily 59 per cent of that of the average month in any one year, the division of the actual daily

average production in that month by the seasonal index results in a figure which might have been expected if seasonal variation had not been present. A similar division of the actual data for all remaining eleven months by their corresponding indexes of seasonal variation eliminates seasonal variation from these monthly production data also. The results appear in Table 130, and graphically in Chart 122. A comparison with Chart 118, which presents the actual daily average production of Portland cement, indicates clearly the course of production after all variations due to seasonal movements have been eliminated. Similar

TABLE 131

PER CENT RELATIVES, ACTUAL ITEMS TO ORDINATES OF TREND, FOR DAILY AVERAGE PRODUCTION OF PORTLAND CEMENT IN THE UNITED STATES, MONTHLY*

Month	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938
January	77	66	66	76	74	62	47	35	20	25	20	22	40	26
February	79	71	66	73	71	66	46	29	20	30	21	23	39	25
March	95	86	92	79	75	82	58	33	24	34	27	33	50	
April	122	106	116	107	106	101	82	39	29	44	40	54	64	
May	132	136	133	133	120	125	98	47	41	55	51	68	69	
June	135	143	141	139	129	128	102	56	53	58	56	72	68	
July	133	140	138	134	128	123	97	52	57	52	50	70	69	
August	139	139	145	143	137	128	94	53	54	50	45	76	70	
September	139	140	142	141	131	119	87	57	58	51	46	77	68	
October	135	135	135	133	123	103	74	53	53	42	46	75	67	
November	118	119	117	118	107	82	58	45	52	38	45	68	56	
December	90	87	94	92	82	60	41	28	23	28	36	54	41	

* Based upon data of Table 122. Straight-line trend fitted to interval 1910–1926.

production series adjusted for seasonal variation[1] only are computed monthly by the Board of Governors of the Federal Reserve System, and the series are combined into its monthly index of industrial production. Each series after adjustment for seasonal variation, however, is expressed as a percentage of its 1923–1925 average.

The monthly data for daily average production of Portland cement continue to include the phenomena secular trend, cyclical variation, and random influences. In many instances it is desirable also to eliminate the secular trend. Such *adjusted relatives*—actual monthly production corrected for both secular trend and

[1] The method employed in the calculation of seasonal indexes differs, however. Instead of using link relatives the board employs the ratios of actual item to the twelve-month moving average centered at the seventh month.

seasonal variation—are computed for many important statistical time series by *The Annalist*, The Federal Reserve Bank of New York, and various other statistical agencies. The elimination of seasonal variation in such cases involves the division of the seasonal index for any month into the per cent ratio of actual item to the ordinate of trend for that month (discussed in Chap. XX). The relatives, actual item to ordinate of trend, for the Portland cement daily average production series, appear in Table 131. The January seasonal index, 59 per cent, is divided into the ratio of the actual to ordinate for every January in the interval; and the February seasonal index, 61 per cent, into the ratio for every February in the series, etc. Thus for January, 1925, the adjusted

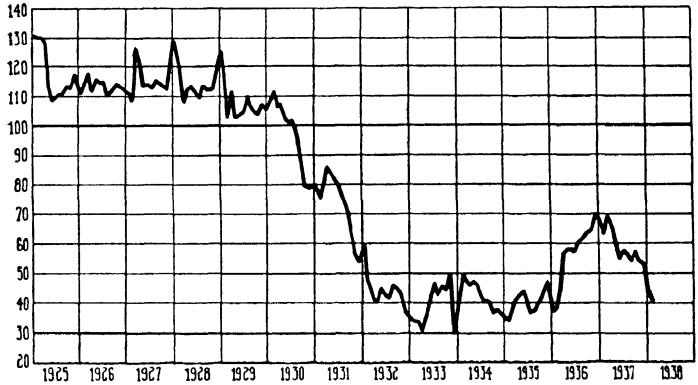


CHART 123.—Daily average production of Portland cement adjusted for secular trend and seasonal variation.

(Data in Table 132.)

relative is 131 (the result of dividing the ratio of actual to ordinate, 77 per cent, by the seasonal index, 59 per cent), and that for February, 1925, is 130. Chart 123 shows the curve of these adjusted relatives, and the fluctuations of this curve represent the cyclical movements (including the irregular, or random, elements, as remarked in Chap. XX) of the Portland cement daily average production series. The extent of the decline from 1929 to 1933, the severity of the depression in the industry from 1932 to 1935, the recovery in 1936, and the renewed decline in 1937 are readily apparent.

The actual work sheet for these calculations can be drawn up as indicated in Table 132, which shows actual items, ordinates of trend, relatives to trend, seasonal indexes, and adjusted relatives in successive columns.

TABLE 132

WORKING TABLE FOR THE DERIVATION OF ADJUSTED RELATIVES FOR DAILY AVERAGE PRODUCTION OF PORTLAND CEMENT

Year and month	Actual item (thousand barrels)	Ordinate of trend (thousand barrels)	Relative, actual to trend (%)	Seasonal index (%)	Adjusted relative
1929					
.....	.....	..	...		...
November	468 4	439 7	107	104	103
December	361 8	440.8	82	77	107
1930					
January	274 1	442 0	62	59	105
February	291 5	443 1	66	61	108
March	362 1	444 3	82	73	112
April	450 7	445 4	101	95	106
May	556 2	446 6	125	117	107
June	574 6	447 7	128	124	103
July	550 8	448 9	123	122	101
August	575 0	450 0	128	125	102
September	537 5	451 2	119	124	96
October	464 9	452 3	103	118	87
November	369 9	453 4	82	104	79
December	273 6	454 6	60	77	78
1931					
January	212 7	455 7	47	59	80
February	211 4	456 9	46	61	75
.....		.. ..	..	..	. .

ANALYSIS OF QUARTERLY DATA

To enable the student to check rather easily the steps involved in the elimination of secular trend and seasonal variation, which have been complicated to some extent by the use of monthly data, a similar analysis for a quarterly series has seemed desirable. The monthly average production by quarters of boots, shoes, and slippers in the United States is presented in Table 133 and is shown graphically on a ratio scale in Chart 124. This chart shows clearly that there is a strong upward movement in this series. Study of the chart also reveals, within each year, a fairly definite seasonal movement: production has a tendency to recede in the second quarter, reach a peak in the third quarter, decline in the fourth quarter, and increase again slightly in the first quarter. Since analysis of the chart indicates the presence of both secular trend and seasonal variation, we may proceed to measure each phenomenon by the methods already described.

TABLE 133

MONTHLY AVERAGE PRODUCTION OF BOOTS, SHOES, AND SLIPPERS IN THE UNITED STATES, QUARTERLY*

Quarter	1922	1923	1924	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937
First quarter	26 34	32 29	27 40	27 47	26 50	27 85	29 38	28 62	27 02	24 41	25 96	25 89	30 71	31 55	33 75	40 95
Second quarter	25.97	30 36	25 23	26 01	24 94	27 17	26 78	28 89	25 80	28 73	24 00	31 82	32 34	31.02	31.01	36 72
Third quarter	26 22	27 61	24 86	27 67	28 79	32 26	31 37	33 83	27 29	31 12	28 37	34 00	30 73	34 47	39.11	35 84
Fourth quarter	29 43	26.78	26 92	26 70	27 95	27.25	27.25	29 13	21.27	21.15	26 11	25.08	25.25	30.87	34 55	23.81

* Unit: million pairs. Compiled from Table 111.

TABLE 134

PER CENT RELATIVES, ACTUAL ITEMS TO ORDINATES OF TREND, FOR MONTHLY AVERAGE PRODUCTION OF BOOTS, SHOES, AND SLIPPERS IN THE UNITED STATES, QUARTERLY*

Quarter	1922	1923	1924	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937
First	100	122	102	101	97	100	105	101	94	84	89	87	103	104	109	133
Second	99	114	94	96	91	98	95	102	90	99	82	107	108	102	101	119
Third	100	104	92	101	104	116	111	119	95	107	96	114	102	114	128	116
Fourth	111	100	100	98	101	97	96	102	74	72	88	84	84	101	112	77

* Based upon data of Table 133. Straight-line trend fitted to the interval 1922–1936.

The quarterly equation of trend for monthly average boot, shoe, and slipper production may be obtained very easily from the annual equation ascertained in Chap. XX. Using the period 1922–1936, the annual trend equation as computed in Table 112 is

$$Y = 28.477 + 0.3077x$$

where Y is the monthly average production of boots and shoes in units of million pairs and x is the number of years measured from July 1, 1929. The annual increment from the average month in one year to the average month in the next year is 307,700 pairs.

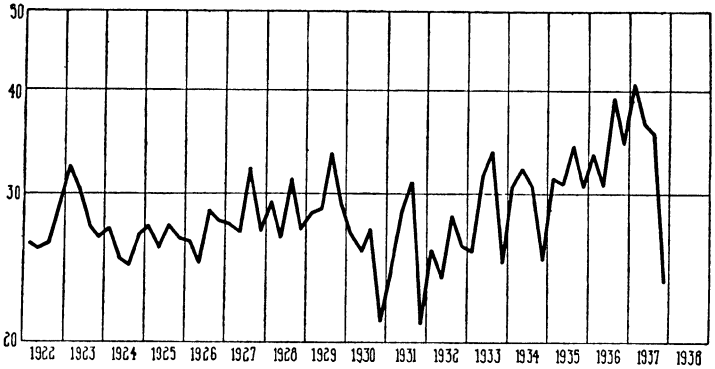


CHART 124.—Monthly average production of boots, shoes, and slippers in the United States, quarterly.
(Unit: million pairs. Data in Table 133. Ratio scale.)

The increment, due to trend, from the average month in one quarter to the average month in the next quarter is one-fourth of the annual increment, or 76,900 pairs, and the quarterly trend equation is therefore

$$Y = 28.477 + 0.0769x$$

where Y is the monthly average production in units of million pairs and x is now the number of quarters from July 1, 1929. Any trend equation, however, should be centered at the middle of the period to which it refers. Since the increase from one quarter to the next is 76,900 pairs, the increase which takes place in a half quarter, from July 1, 1929 to August 15, 1929, is 38,450 pairs and the equation now becomes

$$Y = 28.477 + 0.0384 + 0.0769x$$

or

$$Y = 28.515 + 0.0769x$$

where Y is the monthly average production in units of million pairs and x is the number of quarters from August 15, 1929. From this equation the trend values for each quarter may easily be obtained. The secular trend is eliminated by dividing the actual production in each quarter by the corresponding ordinate of trend for that quarter; these ratios of actual to ordinate are presented in Table 134.

TABLE 135

QUARTERLY LINK RELATIVES FOR BOOT AND SHOE PRODUCTION*

Year	Q_1/Q_4	Q_2/Q_1	Q_3/Q_2	Q_4/Q_3
1922		99	101	112
1923	110	94	91	97
1924	102	92	98	108
1925	102	95	106	96
1926	99	94	115	97
1927	100	98	119	85
1928	108	91	117	87
1929	105	101	117	86
1930	93	95	106	78
1931	115	118	108	68
1932	123	92	118	92
1933	99	123	107	74
1934	122	105	95	82
1935	125	98	111	90
1936	109	92	126	88
1937	119	90	98	66

* Based upon data of Table 133.

The seasonal index is obtained by means of the Persons link relative method. The link relatives of the quarterly averages of monthly boot and shoe production appear in Table 135, and the multiple-frequency table of these links for the period 1923–1930 in Table 136; the period 1931–1937 was not included in the multiple-frequency table because of the erratic movements in production during the recent depression and recovery periods, movements which would have obscured the true seasonal fluctuation. Because of the spread of the links in some quarters, the typical link selected is the average of the middle four items rather than the median. The calculation of the seasonal indexes from these typical links is shown in Table 137, where the logarithmic method of computation is used. The chain index for the fourth quarter (column 6) is 0.981 instead of 1.000, as it should be since the fourth quarter of the preceding year is used as the

base. This discrepancy is caused by the existence of factors other than seasonal variation and to unavoidable inaccuracies in the method, and therefore the chain indexes of each quarter must be adjusted for this discrepancy. The logarithm of the chain index

TABLE 136

MULTIPLE-FREQUENCY DISTRIBUTION OF QUARTERLY LINK RELATIVES FOR MONTHLY AVERAGE BOOT, SHOE, AND SLIPPER PRODUCTION, 1923–1930*

Relatives	$\frac{Q_1}{Q_4}$	$\frac{Q_2}{Q_1}$	$\frac{Q_3}{Q_2}$	$\frac{Q_4}{Q_3}$
Av of middle 4	102½	94½	111	91½
Over 120				
120				
119			|	
118				
117			||	
116				
115			|	
114				
113				
112				
111				
110	|			
109				
108	|			|
107				
106			||	
105	|			
104				
103				
102	||			
101		|		
100	|			
99	||			
98		|	|	
97				||
96				|
95		||		
94		||		
93	|			
92		|		
91		|	|	
90				
89				
88				
87				|
86				|
85				|
84				
83				
82				
81				
80				
Under 80				|

* Based on Table 135.

for the fourth quarter should be 0, and the actual logarithm is 9.9918 − 10, a difference of −0.0082. One-fourth of this discrepancy is applied to the logarithm of the first quarter, two-quarters to the second, etc. (column 7). These corrections added to the logarithms of the chain indexes of column 5 result in the

logarithms of the corrected chain indexes of column 8. The antilogarithms of these indexes appear in column 9. The latter, however, are relative to the preceding fourth quarter as base. They are transferred in column 10 to the average quarter of the year as a base period by expressing each corrected chain index in column 9 as a percentage of 102.3, the average of the corrected chain indexes. The total of the four seasonal indexes equals 400, but in order to obtain this result it was necessary to raise the index for the second quarter from 95.41 to 96; rounding off the indexes to the nearest whole number would have made the index for the

TABLE 137

COMPUTATION OF THE QUARTERLY INDEX OF SEASONAL VARIATION FOR MONTHLY AVERAGE PRODUCTION OF BOOTS AND SHOES IN THE UNITED STATES—LOGARITHMIC METHOD

Quarterly comparison	Link relative		Quarter	Chain index, uncorrected (4th quarter of preceding year = 100)		Correction (applied to logarithm)	Chain index, corrected		Index of seasonal variation (av. for year = 100)
	Actual	Log		Log	Actual value		Log	Actual value	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Q_1/Q_4	1 02½	0 0097	Q_1	0 0097	1 022	0 0020	0 0117	102 7	100
Q_2/Q_1	0 94½	9.9754–10	Q_2	9 9851–10	0 966	0 0041	9 9892–10	97 6	96
Q_3/Q_2	1 11	0 0453	Q_3	0.0304	1 072	0 0062	0 0366	108 8	106
Q_4/Q_3	0 91½	9 9614–10	Q_4	9 9918–10	0 981	0.0082	0 0000	100 0	98
		9 9918–10						409 1	400
								av. = 102 3	

second quarter 95, and the total would have been 399 (see discussion, page 340).

Seasonal variation may be eliminated from the actual monthly average production in each quarter by dividing the actual item in any quarter by its appropriate seasonal index. Seasonal variation may be eliminated from the ratios of actual items to trend by dividing such ratios by the appropriate seasonal indexes, as was done in the case of daily average production of Portland cement, or by the subtraction of the seasonal index from the per cent ratio to the actual item to the ordinate of trend. In using the latter method, that of subtraction, the seasonal index, 100, is subtracted from the ratio of the actual to the ordinate of trend for every first quarter in the interval; and the second quarter seasonal index, 96, from the ratio for every second quarter, and so on. Thus for the first quarter of 1922 the *per cent deviation from the trend corrected for seasonal variation* (the result of correcting for both trend and

seasonal variation) is $100 - 100 = 0$, and that for the second quarter of 1922 is $99 - 96 = 3$. These deviations are the deviations from "normal," if *normal* is defined as a composite of trend and seasonal variation—a definition open to important theoretical qualifications. Since adjusted relatives are usually expressed as a percentage of normal which is always assumed to be 100, 100 should be added to each of these deviations. For the first quarter of 1922 the *adjusted relative* of monthly average production of boots and shoes is accordingly 100 (100 + 0), and for the second quarter 103 (100 + 3). Thus in the second quarter production was 3 per cent above normal. Chart 125 shows the curve of these adjusted relatives, and the fluctuations of this curve

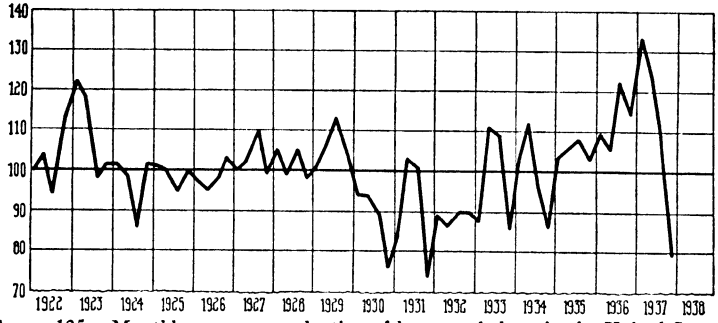


CHART 125.—Monthly average production of boots and shoes in the United States adjusted for secular trend and seasonal variation, quarterly.
(Data in Table 138.)

represent the cyclical movement (including the irregular, or random, elements) of monthly average boot and shoe production by quarters.

The actual work sheet can be drawn up as indicated in Table 138, which shows actual items, ordinates of trend, relatives to trend, seasonal indexes, per cent deviations from normal, and adjusted relatives in successive columns.

The scheme outlined above for eliminating the seasonal variation by the subtraction method involves an indirect assumption. It assumes that the seasonal variation is a fairly uniform percentage of the trend; the subtraction of the seasonal index from the ratio of the actual item to the ordinate of trend really amounts to subtracting from the actual item a percentage (the seasonal per cent) of the ordinate, and dividing the resulting difference by the ordinate. From a practical point of view the adjusted relatives obtained by this method or by the division method described

earlier are approximately the same if the seasonal index does not fluctuate too violently and if the series does not have very wide cyclical fluctuations, or a very steep trend. In the case of quarterly averages of monthly boot, shoe, and slipper production the adjusted relatives computed by either method are the same

TABLE 138

WORKING TABLE FOR THE DERIVATION OF ADJUSTED RELATIVES FOR MONTHLY AVERAGE PRODUCTION OF BOOTS, SHOES, AND SLIPPERS, QUARTERLY

Year and quarter	Actual item (million pairs)	Ordinate of trend (million pairs)	Relative, actual to trend (%)	Seasonal index	Per cent deviation, corrected	Adjusted relative
1928						
Q_3	31 37	28 21	111	106	+ 5	105
Q_4	27 25	28 28	96	98	− 2	98
1929						
Q_1	28 62	28 36	101	100	+ 1	101
Q_2	28 89	28 44	102	96	+ 6	106
Q_3	33 83	28 52	119	106	+13	113
Q_4	29.13	28 59	102	98	+ 4	104
1930						
Q_1	27.02	28 67	94	100	− 6	94
Q_2	25.80	28.75	90	96	− 6	94
...	.....	.....	..	...	. .	...

for most quarters; and, in the few cases for which the adjusted relatives are not the same, the difference is not greater than one. Where the seasonal index does fluctuate over a wide amplitude and where also the cyclical fluctuations are very wide, as in the case of daily average production of Portland cement, the division method —ratio of actual item to ordinate of trend, divided by the appropriate seasonal index—should be used.

CHAPTER XXII

CYCLICAL FLUCTUATIONS

CYCLICAL FLUCTUATIONS IN DIFFERENT SERIES

Processes discussed in the two preceding chapters enable us to determine the secular trend and seasonal variation in a given economic time series and to eliminate the influences of these two types of fluctuation from the original data. The resulting figures appear in the form of adjusted relatives—actual items corrected for secular trend and seasonal variation—as exhibited in Table 132 and in Chart 123, or they may appear in the form of *per cent* deviations from "normal" as exhibited in Table 138; such deviations represent the difference between the actual adjusted relative and normal (100), and may easily be obtained by subtracting 100 from the adjusted relatives. The chart of adjusted relatives, or of per cent deviations from normal, exhibits the same movements; in the case of the former all points on the curve show percentages of normal, and the distances from any point to the 100 line represent deviations from normal; in the case of the latter the distances from the zero line to any point represent deviations from normal. For the pig iron production series of Table 139, these deviations from normal have the values given in Table 140, and are exhibited graphically in Chart 126 for the interval 1925–1938. These deviations reflect the *cyclical fluctuations*, including such irregular elements as exist. It is not feasible to eliminate these irregular fluctuations.

If the object of the investigation were merely to examine this single series, there would probably be no occasion for carrying the computations beyond this point. Table 140 and Chart 126 give an accurate picture of the cyclical movements: they show at what times production of pig iron was increasing, when it reached the peak of boom, when it was declining, and when it was in the trough of depression. Moreover, the results in this form furnish a means of comparing the intensity of the cyclical fluctuations at different times; for example, the magnitude of the movement from 1927 to 1929 can be compared with that from 1933 to 1937.

TABLE 139

DAILY AVERAGE PRODUCTION OF PIG IRON IN THE UNITED STATES, MONTHLY*

	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938
January	108 72	106 97	100 12	92 57	111.04	91 21	55 30	31.38	18 35	39 20	47 66	65.35	103 60	46.10
February	114.79	104 41	105 02	100 00	114.51	101.39	60 95	33 25	19 80	45 13	57.45	62.89	107.12	46.37
March	114 98	111.03	112 37	103 22	119.82	104 72	65 56	31 20	17 48	52 24	57 10	65 82	111 60	46.85
April	108.63	115.00	114 07	106 18	122.09	106 06	67 32	28 43	20 79	57.56	55.45	80.12	113.06	
May	94 54	112 30	109 38	105 93	125.74	104 28	64 32	25 28	28 62	65 90	55 71	85.43	114.10	
June	89.12	107.84	102 99	102.73	123.91	97 80	54 62	20 94	42 17	64.34	51 75	86.21	103.58	
July	85.94	103.98	95 20	99 09	122.10	85 15	47 20	18 46	57 82	39 51	49 04	83.69	112.87	
August	87.24	103.24	95 07	101.18	121.15	81 42	41 31	17 12	59 14	34 01	56 82	87.48	116.32	
September	90 87	104 54	92 50	102 08	116.58	75 89	38 96	19 75	50 74	29 94	59 22	91 01	113 68	
October	97.53	107 55	89 81	108 83	115 74	69 83	37 85	20 80	43 75	30.68	63 82	96 51	93.31	
November	100.77	107 89	88 28	110 08	106 05	62 24	36 78	21 04	36 17	31.90	68 86	98 25	66.89	
December	104.85	99 71	86 96	108.70	91 51	53.73	31.62	17 62	38.13	33.15	67 95	100.48	48.08	

* Unit: thousand long tons. Source: *Iron Age*, monthly.

For purposes involving comparison of the cyclical fluctuations in pig iron production with such fluctuations in other series, however, the per cent deviations from trend corrected for seasonal variation are not entirely satisfactory. The essential difficulty is

TABLE 140

DAILY AVERAGE PIG IRON PRODUCTION IN THE UNITED STATES: PER CENT DEVIATIONS FROM SECULAR TREND CORRECTED FOR SEASONAL VARIATION*

Month	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938
January	11	7	− 2	−11	4	−16	−49	−71	−82	−65	−59	−45	−16	−62
February	14	1	0	− 7	4	− 9	−47	−72	−84	−63	−54	−50	−16	−65
March	9	3	2	− 9	4	−12	−48	−79	−91	−62	−59	−53	−18	−70
April	3	7	4	− 6	7	−10	−46	−80	−87	−57	−60	−41	−16	
May	− 8	7	2	− 3	13	− 8	−45	−80	−78	−47	−56	−33	−12	
June	− 9	7	0	− 2	15	−10	−50	−80	−62	−45	−56	−29	−16	
July	− 9	7	− 4	− 2	17	−18	−53	−78	−45	−61	−54	−27	− 5	
August	− 9	5	− 5	− 1	15	−22	−59	−80	−45	−67	−49	−25	− 4	
September	− 9	2	−11	− 4	7	−31	−65	−82	−56	−74	−51	−26	−10	
October	− 6	2	−17	− 1	3	−40	−69	−84	−65	−77	−50	−25	−29	
November	− 2	3	−18	1	− 5	−46	−69	−83	−71	−75	−45	−23	−49	
December	4	−3	−17	1	−16	−51	−71	−84	−67	−72	−44	−19	−61	

* Secular trend straight line fitted to the interval 1904–1914. Seasonal variation based upon link relatives for the period June, 1903, to August, 1914, and September, 1919, to May, 1923.

that the extent or intensity of the cyclical fluctuation may be greatly different for one series than for another.

Evidence of the seriousness of the difference in extent of fluctuation is obtainable by comparison of curves, on identical time and

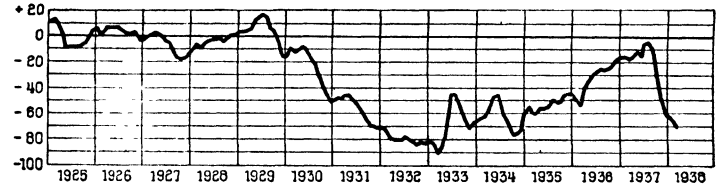


CHART 126.—Per cent deviations from trend corrected for seasonal variation for daily average production of pig iron, monthly.
(Data in Table 140.)

per cent scales, of several series of corrected per cent deviations. One possible procedure consists in plotting two or more such curves on a single chart; but a more flexible plan, quite extensively used in actual practice, places each curve on a single transparent

sheet, and one sheet is superposed on another for comparing the curves. Chart 127 compares the adjusted relatives for daily average boot, shoe, and slipper production and daily average pig iron production in the United States. The actual data for these

TABLE 141

DAILY AVERAGE BOOT, SHOE, AND SLIPPER PRODUCTION IN THE UNITED STATES: PER CENT DEVIATIONS FROM SECULAR TREND CORRECTED FOR SEASONAL VARIATION*

Month	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938
January	3	−2	− 1	2	1	− 3	−24	−17	−15	− 8	3	14	33	−9
February	3	−1	4	7	3	−10	−13	−11	− 6	8	8	12	36	+1
March	1	−4	0	4	3	− 6	− 7	− 8	−15	8	7	5	34	
April	4	−7	0	−5	− 1	− 3	− 1	−14	− 5	15	10	5	26	
May	0	−6	0	−2	6	− 9	6	−16	15	17	7	7	26	
June	−9	−4	4	4	10	− 8	1	−15	22	2	1	1	16	
July	−6	−2	11	9	11	−11	1	−21	25	3	11	18	19	
August	−5	−4	9	7	13	−11	6	−11	8	2	8	23	13	
September	−5	1	8	4	14	−12	− 7	1	− 7	−14	3	20	4	
October	−1	4	5	1	12	−20	−27	2	− 7	−12	1	14	−16	
November	−2	0	− 3	−3	3	−28	−30	−11	−17	−17	− 3	7	−28	
December	1	2	− 2	−5	− 6	−27	−21	−18	−16	− 6	11	20	−21	

* Secular trend straight line fitted to the interval 1922–1936 Seasonal variation based upon link relatives for the period January, 1922, to December, 1930.

charts appear in Tables 140 and 141. The extent of variation in the production of pig iron is much greater than that in the production of boots and shoes.

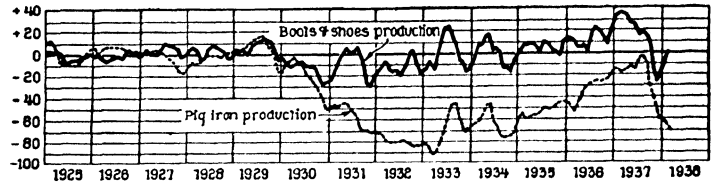


CHART 127.—Comparison of per cent deviations from trend corrected for seasonal variation for daily average production of pig iron and daily average production of boots, shoes, and slippers, monthly.

(Data in Tables 140 and 141)

DIFFERENCES IN FORM AND EXTENT OF CYCLICAL SWINGS

Chart 127 shows that, whereas the timing of the fluctuations is frequently the same for both curves (exceptions may be noted particularly in the period 1931–1934), the form and extent of the cyclical swings differ. The ascent and descent are generally more steep for the iron series than for boots and shoes, and the extent

of the deviation is generally smaller for the latter series than for the former.

Although the form of the fluctuation cannot in practice be treated directly by statistical devices, the extent (called *amplitude*, by analogy to the problem of vibration in physics) is substantially a phenomenon of dispersion. Therefore, to compare the extent of the cyclical fluctuation for two series, the measures of dispersion for their respective corrected per cent deviations are used. Although serious theoretical objections apply to the use of the standard deviation as a measure of dispersion in a time series, search for a better substitute has been inconclusive, and the amplitudes of two series of corrected per cent deviations are customarily compared by means of their standard deviations.

To get the standard deviation of the series in Table 141, the special computation formula developed in Chap. XII need not be used. In the first place, the items are not very numerous and need not be grouped in a frequency table; in getting the standard deviation, it is practicable to square each individual item. Secondly, the items are already in the form of deviations from the mean; for, with a negligible discrepancy due to approximations in the process of correcting for secular trend and seasonal variation, the "trend corrected for seasonal variation" *is* in a sense a mean, and the sum of all the items should be nearly zero. This is not always strictly true unless the standard deviation is computed for the same interval for which the trend was fitted, and even then there may be a slight discrepancy. If different intervals are used, it will probably be necessary to find the mean, M, of the corrected per cent deviations and use M^2 as a correction in the usual manner in finding the standard deviation (as for Table 140).

In the calculation of σ for daily average pig iron production and for daily average boot and shoe production the intervals used in both cases are the same—the period 1922–1936. Since the sum of the per cent deviations from the trend corrected for seasonal variation for the period 1922–1936 in the case of pig iron production was not equal to zero, it was necessary to calculate M as described above to obtain the standard deviation. The result (the square root of the difference between the mean of the squares of the deviations and the square of the average deviation) is 31.27. Similarly, the standard deviation of the per cent deviations from the trend corrected for seasonal variation of daily average boot and shoe production for the period 1922–1936 (Table 141) is found to be 10.72; here, however, it is not necessary to calculate M, since

the standard deviation is based on the same interval as that for secular trend. The amplitude of the per cent deviations from normal, then, is approximately three times as great for pig iron production as for boot and shoe production. That series which has the greater standard deviation has the greater extent of cyclical fluctuation and is, on the whole, the more responsive to the cyclical changes in economic conditions.

CYCLES EXPRESSED IN STANDARD UNITS

The results given in Table 140 can be converted so that they are comparable to similarly converted results for any other series. The data in the table are expressed in units of 1 per cent (based on the normal trend). If each item is divided by the standard deviation, it is said to be expressed in *standard units*. The resulting monthly items are called *cycles*. Thus, the cycle for daily average pig iron production for January, 1925, is 11/31.27, which is 0.4; Table 142 presents the entire series of cycles computed from Table 140. The cycles of daily average production of boots and shoes appear in Table 143.

The effect of this operation of dividing by the standard deviation is to eliminate the confusing difference in extent of fluctuation. Thus, if the intensity of cyclical fluctuation of one series is about three times that of another series, the individual items (per cent deviations, as in Table 140) of the first series tend, on the whole, to be about three times those of the second series, and the standard deviation of the first series is about three times that of the second series. If new series are formed by simply dividing every item of each series by the standard deviation for that series, the effect of "extent of fluctuation" will have been canceled out, and the various resulting series will be comparable. The cycles are the final form to which the figures of the given economic time series are generally reduced. They exhibit the cyclical fluctuations of the series in that form which is best adapted to examination with a view to determining the relation of the particular series to oscillations of prosperity and depression in general business.

Chart 128 compares the cycles of daily average production of pig iron and of boots and shoes. The differences in form and timing of the movements have become more prominent. The graphic comparison, usually with superposed transparent sheets on which the curves of cycles are plotted, is frequently the final step in a study of cyclical movements. The comparison of each curve with the others of the group or with accepted standard cycle

TABLE 142

CYCLES OF DAILY AVERAGE PRODUCTION OF PIG IRON*

Month	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938
January	0 4	0.2	−0.1	−0 4	0 1	−0 5	−1 6	−2 3	−2.6	−2 1	−1 9	−1 4	−0.5	−2.0
February	0 4	0	0	−0 2	0 1	−0 3	−1 5	−2 3	−2 7	−2 0	−1.7	−1 6	−0 5	−2.1
March	0 3	0.1	0 1	−0 3	0 1	−0 4	−1 5	−2 5	−2 9	−2 0	−1 9	−1.7	−0 6	−2.2
April	0 1	0.2	0 1	−0 2	0 2	−0 3	−1 5	−2 6	−2 8	−1.8	−1 9	−1 3	−0 5	
May	−0 3	0 2	0 1	−0 1	0 4	−0 3	−1 4	−2 6	−2 5	−1.5	−1 8	−1 1	−0.4	
June	−0.3	0.2	0	−0.1	0.5	−0 3	−1 6	−2 6	−2 0	−1 4	−1.8	−0.9	−0.5	
July	−0.3	0.2	−0 1	−0.1	0.5	−0 6	−1 7	−2 5	−1 4	−1 9	−1 7	−0 9	−0 2	
August	−0 3	0.2	−0 2	0	0 5	−0 7	−1 9	−2 6	−1 4	−2 1	−1 6	−0 8	−0 1	
September	−0.3	0 1	−0 4	−0 1	0 2	−1 0	−2 1	−2 6	−1.8	−2 4	−1.6	−0 8	−0 3	
October	−0 2	0 1	−0 5	0	0.1	−1 3	−2 2	−2 7	−2 1	−2 5	−1 6	−0 8	−0 9	
November	−0 1	0 1	−0 6	0	−0 2	−1 5	−2 2	−2 7	−2 3	−2 4	−1.4	−0 7	−1.6	
December	0 1	−0 1	−0 5	0	−0 5	−1 6	−2 3	−2 7	−2 2	−2.3	−1 4	−0 6	−1.9	

* Unit: one standard deviation, based upon the interval 1922–1936. Secular trend: straight line fitted to the interval 1904–1914. Seasonal variation: based upon link relatives for the period June, 1903 to August, 1915 and September, 1919 to May, 1923.

TABLE 143

CYCLES OF DAILY AVERAGE BOOT AND SHOE PRODUCTION*

Month	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938
January	0 3	−0.2	−0.1	0 2	0 1	−0 3	−2 2	−1 6	−1 4	−0 7	0 3	1.3	3.1	−0.8
February	0 3	−0 1	0 4	0 7	0 3	−0 9	−1 2	−1 0	−0 6	0 7	0 7	1 1	3.4	0.1
March	0 1	−0 4	0	0 4	0 3	−0 6	−0 7	−0 7	−1 4	0 7	0 7	0 5	3.2	
April	0 4	−0 7	0	−0 5	−0 1	−0 3	−0 1	−1 3	−0 5	1 4	0 9	0.5	2.4	
May	0	−0 6	0	−0 2	0 6	−0 8	0 6	−1 5	1 4	1 6	0 7	0 7	2.4	
June	−0 8	−0 4	0 4	0 4	0 9	−0 7	0 1	−1 4	2 1	0 2	0 1	0 1	1.5	
July	−0 6	−0 2	1.0	0 8	1 0	−1 0	0 1	−2 0	2 3	0 3	1 0	1 7	1.8	
August	−0 5	−0 4	0 8	0 7	1 2	−1 0	0 6	−1 0	0 7	0 2	0 7	2 1	1 2	
September	−0.5	0 1	0 7	0 4	1 3	−1 1	−0 7	0 1	−0 7	−1 3	0 3	1.9	0 4	
October	−0 1	0 4	0 5	0 1	1 1	−1 9	−2 5	0 2	−0 7	−1 1	0 1	1 3	−1 5	
November	−0 2	0	−0 3	−0 3	0 3	−2 6	−2 8	−1 0	−1 6	−1 6	−0 3	0 7	−2.6	
December	0 1	0 2	−0 2	−0 5	−0 6	−2 5	−2 0	−1 7	−1 5	−0 6	1 0	1 9	−2 0	

* Unit: one standard deviation, based upon the interval 1922–1936. Secular trend fitted to the interval 1922–1936. Seasonal variation: based upon link relatives for the period January, 1922, to December, 1930.

curves yields highly significant information concerning the response of a particular economic factor to the forces of business prosperity and depression. Here the rapid cyclical fluctuations in boot and shoe production are clearly apparent; this series, a consumers' good industry, fluctuates more rapidly in the cycle than does pig iron production, primarily a capital goods industry.

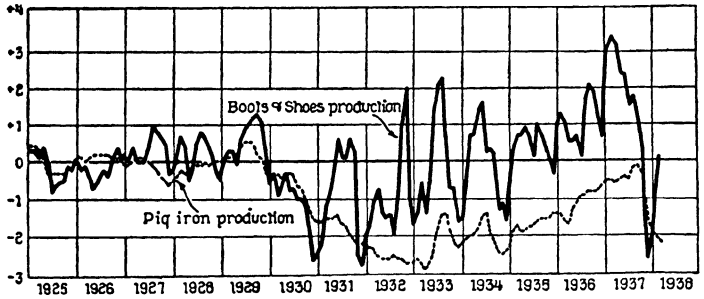


CHART 128.—Comparison of cycles for daily average production of pig iron and daily average production of boots, shoes, and slippers, monthly.
(Unit: one standard deviation. Data in Tables 142 and 143.)

THE BUSINESS CYCLE

The term *cycle* is used also in a more general sense as pertaining to a curve comprising the succession of all monthly or quarterly "cycles" (in the above sense) which constitute a complete wave length in the fluctuation, extending from one peak in the oscillation to the next peak; and it is in this sense that the term finds its wide use in references to the *business cycle.*

The study of any particular series of cycles reveals certain interesting properties. Although the fluctuations occur in an orderly succession of rise, peak, fall, and trough, there is no mathematical regularity about the wavelike motion. One business cycle may be about three years in length, another nearly five; the peak in one cycle may be twice as high as in the succeeding cycle; while in one cycle the rise may be gradual and the subsequent decline very sharp, in another the rise may be vigorous and the decline less precipitous. This lack of uniformity, in respect to length, extent, and shape, of the cyclical movement is the effective obstacle to the application to economic problems of those precise mathematical methods of analysis which are so useful in the study of vibratory motions in physics.

Economic fluctuations, although recurrent, are not periodic in the mathematical sense. Therefore it is impossible, by an

analysis of the cyclical fluctuations of a single series, to forecast (even for the immediate future) the items of that series by considerations of the length, amplitude, or form of the cycles of that series. The essential fact is that when future data are at hand the coming cycles may prove to be markedly different in length or form from those already found in the series.

An analysis of the cycle data for several statistical time series, however, gives the economist a knowledge of the fundamental differences which exist among the cyclical movements of a wide variety of economic factors, and furnishes him with more detailed information than he otherwise would have of the behavior of different series in our dynamic economic society. Comparison of the cycles of several series will indicate, for example, great differences in amplitude and in timing. It is on the basis of such information that the economist will perfect his knowledge of economic movements and improve his understanding of business cycle theory, both of which are necessary for successful interpretation and forecasting of the cycle. Various mechanical methods of forecasting also result from such analysis of the cycle data for several statistical series. Comparison of the cyclical movements for a group of series frequently indicates a regularity in the sequence in the fluctuations of the cycle data for members of the group which have cycles similar in form but which occur successively in time. If it can be shown that one of two series regularly reaches its peak about four months, for example, after the other, and that this same sequence of movements holds true for the trough and the other phases of the cycle, the movements of the second series may be accepted as the basis of probable forecast for the movements of the first. The length of time in months by which the first curve follows (or precedes) the second is called the *lag* (or *lead*) of the first series relative to the second. The statistical verification of such sequence can be ascertained very easily from an analysis of the charts of the cycles of several series, and it may be measured mathematically through the results obtained in the correlation of these time series (see Chap. XXIII). The existence of such sequences is presumably due to fundamental factors and relations which have tended to persist over a long period of time in our economic system. Obviously any change in these fundamental relations may change the nature of the sequences. The successful use of such sequences in forecasting, then, involves an understanding of these relations, which in general must be based on nonstatistical information.

CHAPTER XXIII

LAGGING CORRELATION AND FORECASTING SEQUENCE

CORRELATION WHEN NO LAG EXISTS

In measuring the correlation between time series the technical procedure is somewhat more complicated than the procedure in the case of measuring correlation between frequency series discussed in Chap. XVI. It will be remembered that correlation is the mutual relation which exists between two series of paired variates and that the purpose of our correlation measurements is to ascertain the probable variation which would take place in one series given a certain variation in the other. The time series, however, is itself a composite of variables. The value of any series at any particular time is made up of variations which are caused by the presence of secular trend, seasonal variation, cyclical and irregular variations. Correlation of the actual items of two time series may be caused by the existence of a strong mutual relation between the secular trends of the two series or between the seasonal variations or between the cyclical fluctuations. In some cases no correlation at all will exist between the actual items and yet a substantial degree of correlation may exist between the cyclical data.

Correlation of the actual items of two time series, then, gives very unsatisfactory results. By the correlation process we hope to secure measurements of the mutual relation which exists between two variable phenomena affected by the same causes. To obtain such a relation between the cyclical elements of two time series, it is necessary to eliminate the secular trends and seasonal fluctuations from the actual items and compare only the cyclical fluctuations. The visual comparison of two curves of cycles furnishes estimates of the extent of the correlation between the two series, and of the amount of lag.

The *ordering* of the items in time (the succession of one item after, and its consequent partial dependence upon, another) introduces a complication which was not considered in the original development of the correlation method. Therefore, the same conclusions cannot be drawn from correlation coefficients, and

particularly high coefficients, in time series which can be drawn from similar coefficients for ordinary frequency series. Moreover, it is impossible to emphasize too strongly the necessity of making the visual estimate from the charts as a check upon the numerical result obtained by use of the correlation formula. In some instances the formula may yield a high value of the coefficient, whereas in fact the charts show very slight similarity between the two series.

If there is no lag, the calculation of the correlation coefficient is extremely simple. For example, suppose the correlation is desired between daily average pig iron production and daily average boot and shoe production for the interval 1922–1936. The coefficient of correlation is merely $\Sigma xy/n$, where x is the pig iron cycle and y is the boot and shoe production cycle for a particular month. One multiplies the January, 1925, pig iron cycle, 0.4, by that for boot and shoe production, 0.3, to get 0.12; then multiplies together the February, 1925, cycles, 0.4 and 0.3, to get 0.12; and, in like manner, finds the product of the two cycles for every month in the interval 1922–1936. The sum of these products, divided by 180, is r. The procedure here is simpler than that in Chap. XVI because the x and y, by the very definition of *cycle*, have been divided by their respective standard deviations. In case the trends were not fitted for the interval for which r is to be computed, or if the standard deviations were not found for that interval, the more complicated formula for r shown in Chap. XVI must be used.

The value of r found for the above illustration is 0.40, showing a relatively low degree of correlation, as one would expect from examining Chart 128. It is obvious that the correlation can be measured for each of a large number of pairs of series for which there is no lag, and that an order of similarity can be established: the pair having the highest r has greater similarity between its members than pairs with lower values of r. Small differences in r should not be regarded as significant in making such tests, for the small total frequency (180 here) of the "observations" and the fact that these are time series operate to reduce the precision of r as a measure of similarity. With due precaution, those series which are most closely related to each other or to some accepted standard can be selected from a group of series in which the cyclical fluctuations are synchronous—among which there is no lag.

By such a process of selection, the components of an index of trade or industrial conditions—or other special-purpose index—

can be chosen. The fact that the individual components are highly correlated in pairs indicates that they all may reflect a single fundamental fluctuating phenomenon. The averaging together, into an index, of such components tends to cancel out the individual peculiarities of each component, peculiarities which may be regarded as "errors of observation" or "of sampling" when that component is considered as a measure of the fundamental phenomenon. For example, if there are a priori reasons for believing that pig iron production and boot and shoe production reflect general industrial conditions, and if actual examination shows high correlation and no lag between these series, they may fairly be averaged together to yield an index of industrial conditions. Unless there were adequate reason for using weights, the process would consist merely in adding the two cycles, one for each of the two series, and dividing the total by 2 to yield the index for a particular month. Of course, a good index would probably be based upon more than two series.

NUMERICAL DETERMINATION OF LAG

When the correlation is desired between two series for which a lag exists, as shown by comparison of the curves, difficulties arise. The essential fact is that the lag is not an invariable length of time; in one part of the chart it may appear longer than in another part. Although a visual estimate of the average lag is possible, the result is always in doubt, and one of the chief services of the numerical evaluation of correlation is the determination of that lag which yields maximum correlation.

Chart 129 presents the cycles for pig iron production and stock prices for the period 1919–1930. Given the series of cycles for pig iron, x, and stock prices, z, the correlation is computed for several assumed lags, including that suggested by the chart. For example, correlations should be computed on assumed lags of 1, 2, 3, 4, 5, 6 months: multiplying the February, 1919, x by the January, 1919, z, the March, 1919, x by the February, 1919, z, etc., for the assumed lag of 1 month; the March, 1919, x by the January, 1919, z, the April, 1919, x by the February, 1919, z, etc., for the assumed lag of 2 months; and similarly for the other assumed lags. A summary of the approximate results for the interval 1919–1930 is shown in Table 144. The values of r given in the last column differ for different assumed lags, and the maximum value occurs for a lag of two months. Thus, the method of correlation yields an arithmetic determination of the amount of

lag. There is some tendency during the period 1919–1930 for movements in stock prices to precede movements in pig iron production, but the relation is not very marked.

The bearing of the time interval used upon the degree of correlation and upon the indicated length of lag should not be

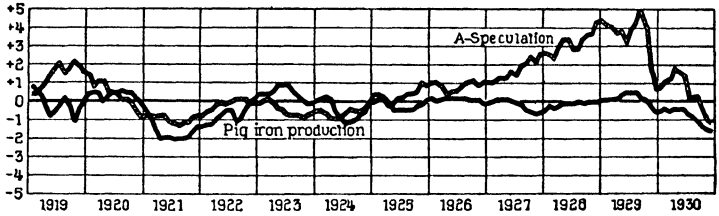


CHART 129.—Cycles of stock prices (*A*) and daily average production of pig iron, monthly

(Unit: one standard deviation. Data in Table 142, page 359, and Tables *DD* and *EE*, pages 388–390.)

overlooked. In the above problem the interval 1919–1930 is a much more homogeneous period than 1929–1936 would be. On the whole it is unwise to study correlation over a time interval in which the relations between the series undergo considerable changes. It should be determined from the original graphic

TABLE 144

CORRELATION OF CYCLES OF DAILY AVERAGE PRODUCTION OF PIG IRON WITH CYCLES OF STOCK PRICES, FOR DIFFERENT ASSUMED LAGS OF PIG IRON PRODUCTION AFTER STOCK PRICES, 1919–1930*

Lag (in months)	Number of pairs of months	Σxz	r
0	144	21 1	0 352
1	143	21 7	0.370
2	142	20 4	0.372
3	141	17.6	0.358
4	140	16 4	0.348
5	139	14 9	0.335
6	138	11 7	0.311

* Based upon data in Table 142 and Tables *DD* and *EE*.

inspection, or by actual trial of a portion of the interval, whether the whole interval can be treated as a unit. For example, the coming of the World War upset the close relations which had previously existed between series; and while it did not ordinarily obliterate the sequence of the economic movements, it worked such temporary changes that the four or five wartime years

should, in most cases, be omitted from studies of this sort. The extent of the recent great depression and the attempts by government to increase business activity through government spending also upset the close relations which had previously existed between series, and therefore this period also should be omitted from such correlation studies. The values of r based on varying periods differ greatly, and the interval selected should be as nearly as possible homogeneous.

THE FORECASTING SEQUENCE

By the process of correlation, then, it is possible (1) to measure the amount of lag between series and thus to classify large numbers of series into groups according to the timing of their cyclical swings, and (2) to compare the degree of similarity for various pairs within any time group, and thus select from the group those which have the greatest *similarity* to each other in their cyclical movements. In this manner several indexes, after the plan of that discussed above, have been constructed, and a striking feature of these indexes is that the cyclical movements in one index lag after the similar cyclical movements in another. The second then *forecasts* by its movements the swings of the first.

Although the theory of the sequence of fluctuations in economic factors was already vaguely held and loosely stated, the scientific verification of this theory in accordance with the principle and by the method of lagging correlation was first accomplished by Professor Warren M. Persons,[1] and he subsequently used the method for the establishment of a scheme of forecasting the movement of particular economic factors, which was satisfactory for economic conditions substantially comparable to those of the prewar test period. This method of forecasting depends upon the assumption that the cycle of the economic factor under consideration will maintain its "customary" relation to the cycles of the other factors. He analyzed, for the prewar interval, 1903–1914, twenty-four series of fundamental importance in the economic life of the nation, and examined the lag and correlation of each with all the others. By means of the differences in lag, the series could be classified into groups such that, within any one group, the lag between the several series was slight and the degree of correlation was high. By a process of selection based partly upon statistical and partly upon practical considerations, some of the original

[1] *Review of Economic Statistics*, Prel. Vol. I, 1919.

twenty-four series were eliminated; and the remaining series fell into three groups: *A, Speculation; B, Business; and C, Money.* These groups have been revised from time to time, and the present data for each group are shown in Chart 130.[1]

THE NATURE OF THE SEQUENCE

The relations between the curves of Chart 130 form the *forecasting sequence* by means of which future movements in one of the three curves can be anticipated by an examination of the recorded movements of the other two. In order to use the sequence successfully for forecasting, it is necessary to get clearly in mind the relations which hold between the heights and directions of movement of the three curves in every phase of the busi-

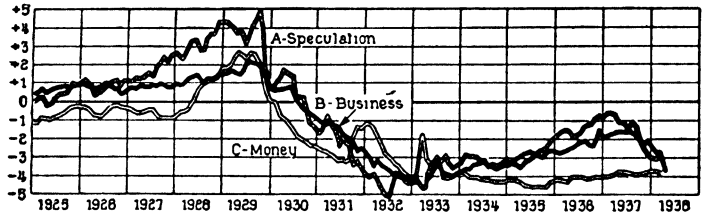


CHART 130.—The Harvard Index of General Economic Conditions
(Unit: one standard deviation. Data in Table *EE*, pages 388–390.

ness cycle: the characteristic relations of the period of depression, of the period of recovery, of the period of boom, and of the period of decline. Moreover, the basic assumption, that current conditions are fundamentally analogous to those of the test period, must not be ignored.

There are intervals when the prospect is not clear-cut; and, in general, for such intervals one of the curves is temporarily out of its usual position relative to the other two. Such doubtful intervals exist in consequence of the fact that the interrelations of the series are not rigidly perfect. The occurrence of such intervals requires the forecaster to avoid using the device mechanically: he must rather use the curves as a guide and look behind the statistical results at any given moment to assure himself that there are no factors present in the situation which might upset the normal sequence of movements. When used with such precaution as is necessary in all statistical treatment of economic problems, the

[1] Edwin Frickey, "Revision of the Index of General Business Conditions," *Review of Economic Statistics*, May 15, 1932, pp. 85–87.

sequence of curves furnishes an analysis of business conditions on the basis of which developments in the immediate future may be predicted—not with confidence and certainty, but much more effectively than would otherwise be possible.

The gradual changes in economic organization may work to modify the forecasting sequence as determined for any particular interval. Although revolutionary rearrangements of the sequence of movements are not likely to occur, except for a short period of time, some considerable modification of the average lags between the curves may well develop. Moreover, somewhat different selections of the constituents of the curves, or indeed certain entirely new index curves, may furnish a more satisfactory sequence for the study of cyclical movements under altered conditions. The principles, however, remain the same: The length and amplitude and form of the cycle are highly variable, but the sequence between the movements of properly classified indexes is relatively stable and changes slowly in the long run.

APPENDIX A

NUMERICAL DATA FOR CERTAIN OF THE CHARTS OF PARTS I TO III

The data upon which many of the charts in the text are based are given in tabulations in earlier pages. For a large number of the charts, however, such tabular presentations are not essential to the textual development; and, with a few exceptions, they are given herewith to facilitate verification of the several charts and to provide material for further study by graphical or analytical methods.

TABLE A

GROSS INCOME OF SUBGROUPS IN SELECTED GROUPS OF MANUFACTURING CORPORATIONS IN 1934*

Food and kindred products	
Bakery and confectionery products	1,212
Canned products	751
Mill products	1,025
Packing-house products	3,116
Sugar	579
Other food products	1,661
Textiles and their products	
Cotton goods	982
Woolen and worsted goods	377
Silk and rayon goods	478
Other textiles	1,068
Clothing	1,797
Knit goods	547
Metal and its products	
Iron and steel	2,167
Locomotives, etc.	143
Machinery	2,031
Motor vehicles	2,946
Household equipment	385
Office equipment	229
Building material and hardware	1,085
Precious metals, etc.	265
Other metal industries	1,377

* Unit: million dollars. Source: "Statistics of Income, 1934, Part 2," Washington, U. S. Treasury, 1937, pp. 60–63.

TABLE B

UNITED STATES TREASURY BONDS, OF OVER TEN YEARS STATED MINIMUM MATURITY, OUTSTANDING ON DECEMBER 31, 1935*

Rate and dates of maturity	Amount
4¼, 1947–1952	759
3¾, 1946–1956	489
3⅛, 1946–1949	819
3 , 1951–1955	755
3 , 1946–1948	1,036
3⅛, 1949–1952	491
2⅞, 1955–1960	2,611

* Unit: million dollars. Source: "Statistical Abstract of the United States, 1936," Washington, U. S. Department of Commerce, 1936, p. 201.

TABLE C

PERCENTAGE OF NET TO GROSS TON-MILES (EXCLUDING LOCOMOTIVES) FOR UNITED STATES CLASS I RAILROADS IN 1936*

Region	%
New England	37 4
Great Lakes	40 8
Central Eastern	46 3
Pocahontas	53 9
Southern	41 1
Northwestern	40 8
Central Western	36 3
Southwestern	37 2
Entire United States	41 6

* Source: "Statistics of Railways in the United States, 1936," Washington, Interstate Commerce Commission, 1937, p. S 154. (Data pertain to freight train performance.)

TABLE D

UNITED STATES NET IMPORTS OF GOLD IN 1937 FROM COUNTRIES MAKING PRINCIPAL SHIPMENTS*

Belgium	90 9
France*	– 13 7
Switzerland	54 5
United Kingdom	891 5
Canada	111 5
Mexico	38.5
Australia	34.7
British India	50 8
Japan	246.5

* Unit: million dollars. Source: *Federal Reserve Bulletin*, March, 1938, p. 206.
* Net export to France.

TABLE E

UNITED STATES INTERNAL REVENUE COLLECTIONS FOR FISCAL YEAR ENDING JUNE 30, 1937*

Individual income tax	1,091 7
Corporation income and excess-profits taxes	1,082 0
Unjust enrichment tax	6 1
Dividend tax	a
Alcohol taxes	594 2
Estate and gift taxes	305 5
Tobacco manufacturers tax	552 3
Sales taxes	617.4
Other miscellaneous taxes	138 1
Social security taxes	265 7
Total collections	4,653 2

* Unit: million dollars. Source "Annual Report of the Commissioner of Internal Revenue, 1937," Washington, 1937, pp 1–2.
[a] Less than 1/10 million dollars.

TABLE F

PERCENTAGE DISTRIBUTION OF NATIONAL INCOME OF UNITED STATES IN 1935 ACCORDING TO INDUSTRIAL ORIGIN*

Manufacturing	21 7
Service	15 6
Government	14 5
Trade	12 9
Transportation and other public utilities	9 4
Agriculture	8 8
Finance	8 7
Mining	2 1
Construction	1 7
Miscellaneous	4 5

* Source Simon Kuznets, "National Income and Capital Formation," 1919–1935, New York, National Bureau of Economic Research, 1937, p 17.

TABLE G

PERCENTAGE WEIGHTS, BASED ON 1926 VALUES, OF SUBGROUPS IN THE FOODS GROUP OF THE U. S. B. L. S. WHOLESALE PRICE INDEX*

Subgroup	%	Fractions of 360°
Dairy products	18 08	65.1
Cereal products	16 61	59 8
Fruits and vegetables	10.91	39 3
Meats	31.23	112 4
Other foods	23 17	83.4

* Source: "Wholesale Prices, 1931," Washington, U. S. Bureau of Labor Statistics, Bulletin 572, January, 1933, pp. 93–96.

TABLE H

ADJUSTED DEMAND DEPOSITS OF ALL MEMBER BANKS OF THE FEDERAL RESERVE SYSTEM ON DATES OF CALL IN 1934*

March 5	3,648
June 30	3,792
October 17	4,168
December 31	4,292

* Unit: million dollars. Source: "Twenty-Third Annual Report of the Board of Governors of the Federal Reserve System," Washington, 1937, p. 142. (Demand deposits other than interbank and U. S. Government, less cash items reported as in process of collection, and less cash items reported on hand but not in process of collection.)

TABLE I

SILVER PRODUCTION FROM MINES IN CONTINENTAL UNITED STATES, ALASKA, PUERTO RICO, AND PHILIPPINE ISLANDS, 1936*

Alaska	398	New Mexico	1,244
Arizona	8,556	Oregon	103
California	2,037	South Dakota	145
Colorado	6,391	Texas	1,348
Idaho	14,815	Utah	11,204
Michigan	...	Washington	60
Montana	11,498	Philippine Islands	461
Nevada	5,173	Other	380

* Unit: thousand fine ounces Source: "Statistical Abstract of the United States, 1937," Washington, U. S. Department of Commerce, 1938, p. 718.

TABLE J

MISCELLANEOUS MONTHLY SERIES FOR 1936 AND 1937*

Month	a.	b.	c.	d.	e.	f.	g.		h.	i.	
							1936	1937		(1)	(2)
January	614	0	811	130	631	72	98	10 18	11.36	1245	5626	15928
February	521	6	721	119	553	76	89	10 17	11.44	1282	5617	14244
March	382	3	820	165	754	90	108	10 18	11 57	1240	7616	19823
April	317	1	786	195	708	89	123	10 22	11.80	1181	7176	20230
May	261	5	834	151	552	95	98	10.40	11.99	1030	8614	22820
June	560.	3	670	180	1300	90	69	10.61	12.32	898	8463	22254
July	340	2	618	191	659	65	91	10 65	12 45	736	7706	20409
August	187	3	707	178	556	72	96	10 72	12.57	582	6780	19761
September	223	8	564	127	680	100	45	10 84	12.74	440	7819	24806
October	197	5	768	124	616	103	73	11 04	12 80	394	8957	29990
November	135	3	786	106	542	101	101	11.18	12 77	447	8373	27095
December	163	8	932	94	730	156	142	11 26	12 76	582	16615	38005

* Series (by columns), units, and sources are as follows:

a. Volume of new financing in the United States, 1937. Unit: million dollars. Source: *Commercial and Financial Chronicle*, January 8, 1938, p. 167.

b. Total business failures, 1937. Unit: one failure. Source: *Survey of Current Business*, February, 1938, p. 32.

c. Construction contracts awarded, privately financed, 1937. Unit: million dollars. Source: *Federal Reserve Bulletin*, February, 1938, p. 159.

d. Total expenditures of U. S Treasury, 1937. Unit: million dollars. Source. *Federal Reserve Bulletin*, February, 1938, p. 149.

e. Department store sales, unadjusted for seasonal variation, 1937. Unit: per cent of 1923–1925 average. Source: *Federal Reserve Bulletin*, February, 1938, p. 152.

f. Sugar meltings, adjusted for seasonal variation, 1937. Unit: per cent of 1923–1925 average. Source. *Federal Reserve Bulletin*, February, 1938, p. 153.

g. Monetary gold stock of United States at end of each month, 1936–1937. Unit: billion dollars. Source: *Federal Reserve Bulletin*, February, 1938, p. 136. Figure for December 31, 1935 is 10.12.

h. Stocks of meats, cold storage, at end of each month, 1937. Unit: million pounds. Source: *Survey of Current Business*, February, 1938, p. 43. Figure for December 31, 1936 is 1149.

i. Sales, two chain-store systems, 1937. (1) W. T. Grant & Co., (2) J. C. Penney Co. Unit: thousand dollars. Source: *Survey of Current Business*, February, 1938, p. 27.

TABLE K

STATUTORY NET INCOME LESS DEFICIT OF ALL CORPORATIONS IN THE UNITED STATES, 1924–1935*

1924	5,363	1930	1,551
1925	7,621	1931	−3,288
1926	7,505	1932	−5,644
1927	6,510	1933	−2,547
1928	8,227	1934	94
1929	8,740	1935	1,696

* Unit: million dollars. Source: "Statistics of Income, 1934, Part 2," Washington, U. S. Treasury, 1937, p. 36. 1935 figure is from preliminary copy of source.

TABLE L

GROSS INCOME OF ALL CORPORATIONS IN THE UNITED STATES, 1919–1933*

1919	99 9	1927	144.4
1920	118 2	1928	152 8
1921	91 2	1929	160 6
1922	100 9	1930	136.1
1923	118 6	1931	107 5
1924	119 2	1932	81 1
1925	134 3	1933	83 6
1926	142 1		

* Unit: billion dollars. Source: "Statistics of Income, 1934, Part 2," Washington, U. S. Treasury, 1937, p. 35.

TABLE M

EXCESS RESERVES OF MEMBER BANKS OF THE FEDERAL RESERVE SYSTEM, 1937*

January 31	2,152	July 31	791
February 27	2,078	August 31	773
March 31	1,398	September 30	1,038
April 30	1,594	October 30	1,055
May 31	918	November 30	1,169
June 30	865	December 31	1.212

* Unit: million dollars. Source: *Federal Reserve Bulletin*, February, 1938, p 128. December 31, 1936 figure is 1,984.

TABLE N

ISSUES OF CORPORATE LONG-TERM BONDS, FOR NEW CAPITAL, IN THE UNITED STATES, 1919–1937*

1919	499	1929	1,873
1920	1,039	1930	2,460
1921	1,275	1931	951
1922	1,537	1932	271
1923	1,833	1933	24
1924	1,924	1934	112
1925	2,231	1935	323
1926	2,418	1936	816
1927	2,962	1937	743
1928	2,175		

* Unit: million dollars. Source: *Survey of Current Business*, February, 1938, pp. 14–20.

TABLE O

YIELD PER ACRE OF COTTON IN THE UNITED STATES, BY STATES, 1936*

Alabama	236	New Mexico	457
Arizona	438	North Carolina	298
Arkansas	227	Oklahoma	62
California	574	South Carolina	279
Florida	170	Tennessee	250
Georgia	228	Texas	121
Louisiana	260	Virginia	298
Mississippi	305	Other states	313
Missouri	360		

* Unit: one pound Source. "Statistical Abstract of the United States, 1937," Washington, U. S. Department of Commerce, 1938, pp. 653–655.

TABLE P

NUMBER OF CONCERNS IN BUSINESS IN THE UNITED STATES, BY STATES, 1936*

Maine	1 50	West Virginia	2 04
New Hampshire	0 94	North Carolina	3.08
Vermont	0 65	South Carolina	1.46
Massachusetts	7 85	Georgia	3 03
Rhode Island	1 19	Florida	2 68
Connecticut	2 91		
		Kentucky	3 05
New York	24 54	Tennessee	2 78
New Jersey	8 48	Alabama	2 13
Pennsylvania	16 33	Mississippi	1 71
Ohio	11 13	Arkansas	2 16
Indiana	5 35	Louisiana	2 46
Illinois	13 41	Oklahoma	3 40
Michigan	6 77	Texas	9 20
Wisconsin	5 77		
		Montana	0 93
Minnesota	4 62	Idaho	0 76
Iowa	4 59	Wyoming	0 41
Missouri	6 48	Colorado	1 99
North Dakota	1 05	New Mexico	0 61
South Dakota	1 22	Arizona	0 55
Nebraska	2 71	Utah	0 76
Kansas	3.55	Nevada	0 21
Delaware	0 45	Washington	3 40
Maryland	2 83	Oregon	2 02
District of Columbia	0 86	California	12 09
Virginia	2 93		

* Unit: 10,000 concerns Source: "Statistical Abstract of the United States, 1937," Washington, U. S. Department of Commerce, 1938, p. 294.

TABLE Q

NUMBER OF NATIONAL BANKS IN OHIO (EXCLUDING "RESERVE CITIES") HAVING TOTAL DEPOSITS WITHIN SPECIFIED RANGES ON DECEMBER 31, 1936*

Class interval[a]	Number of banks	Deposits[b]
0- 500	50	16,647
500- 1,000	71	52,828
1,000- 1,500	33	40,915
1,500- 2,000	28	48,062
2,000- 2,500	12	27,645
2,500- 3,000	10	27,285
3,000- 3,500	8	25,214
3,500- 4,000	4	15,098
4,000- 4,500	1	4,222
4,500- 5,000	2	9,463
5,000- 5,500	3	15,420
5,500- 6,000	3	17,185
6,000- 6,500	1	6,431
6,500- 7,000	3	20,224
7,000- 7,500	0	0
7,500- 8,000	2	15,310
8,000- 8,500	2	16,125
8,500- 9,000	0	0
9,000- 9,500	0	0
9,500-10,000	2	19,647
.....	.....	.....
13,500-14,000	1	13,938
.....	.....	.....
20,000-20,500	1	20,174
20,500-21,000	1	20,753
.....	.....	.....
over 21,000	1	33,392

* Based upon data from "Individual Statements of Condition of National Banks at Close of Business, December 31, 1936" (Table N, issued as supplement to Annual Report of Comptroller of the Currency), Washington, 1937, pp. 121–129. Four banks in Cincinnati, two in Cleveland, three in Columbus, and one in Toledo (the "reserve cities") have been omitted.

[a] Lower limit inclusive. Unit: thousand dollars.

[b] Unit: thousand dollars.

TABLE R

NUMBER OF NATIONAL BANKS IN MICHIGAN (EXCLUDING "RESERVE CITIES") HAVING TOTAL DEPOSITS WITHIN SPECIFIED RANGES ON DECEMBER 31, 1936*

Deposits[a]	Frequency
0– 500	13
500– 1000	22
1000– 1500	9
1500– 2000	4
2000– 2500	6
2500– 3000	4
3000– 3500	3
3500– 4000	2
4000– 4500	0
4500– 5000	0
5000– 5500	3
5500– 6000	2
6000– 7000	3
7000– 8000	2
8000–10000	2
10000–12000	1
12000–14000	2
..	..
26000–28000	1
Total	79

* Compiled from same source and for same banks as described in Table 61.
[a] Upper limit excluded. Unit: thousand dollars.

TABLE S

NUMBER OF NATIONAL BANKS IN MICHIGAN (EXCLUDING "RESERVE CITIES") HAVING LOANS WITHIN SPECIFIED RANGES ON DECEMBER 31, 1936*

Loans[a]	Frequency	Loans[a]	Frequency
0	12	15	1
1	16	16	1
2	14	17	3
3	4	18	0
4	6	19	1
5	2	20	0
6	3	21	1
7	2	..	..
8	2	26	1
9	2	..	..
10	4	38	1
11	0	..	..
12	0	51	1
13	0	Total	79
14	2		

* Compiled from same source and for same banks as described in Table 61.
[a] Lower limit, inclusive, of class intervals $100,000 wide. Unit: $100,000.

TABLE T

NUMBER OF FEMALE "ROVING-FRAME TENDERS" OVER 16 YEARS OF AGE EMPLOYED IN THE COTTON MILLS OF THE UNITED STATES, RECEIVING SPECIFIED WEEKLY WAGES IN 1890*

Wages[a]	Actual series	Samples				
		(1)	(2)	(3)	(4)	(5)
$2 50	2				1	
3 00	8		1	1	2	
3 50	13	1	1	2	3	3
4 00	9	1	0	0	4	4
4 50	34	2	9	6	11	5
5 00	50	12	18	13	14	14
5 50	52	18	20	12	17	6
6 00	103	27	24	20	19	33
6 50	68	19	16	29	19	16
7 00	31	10	7	10	3	11
7 50	11	5	2	3	1	2
8 00	17	5	2	4	6	6
8 50	1					
9 00	1					
Total	400	100	100	100	100	100
Mean	6 03	6 34	6 00	6 26	5 87	6 19
σ	1 11	.87	.87	95	1 13	1 02

* Source: "Twelfth Census of the United States, 1900, Special Report on Employees and Wages," Washington, U. S Census, 1903, p 31.
[a] Lower limit, inclusive, of class interval Unit, one dollar.

TABLE U

NUMBER OF NATIONAL BANKS IN MICHIGAN (EXCLUDING "RESERVE CITIES") HAVING UNITED STATES SECURITY HOLDINGS WITHIN SPECIFIED RANGES ON DECEMBER 31, 1936*

Holdings[a]	Frequency	Holdings[a]	Frequency
0	20	18	0
1	13	19	0
2	10	20	1
3	5	21	3
4	6	22	1
5	5	23	0
6	1	24	1
7	0	.	.
8	0	27	1
9	2	.	.
10	3	37	1
11	0	..	..
12	1	40	1
13	1	..	..
14	0	60	1
15	0	..	.
16	1	73	1
17	0	Total	79

* Compiled from same source and for same banks as described in Table 61.
[a] Lower limit, inclusive, of class intervals $100,000 wide. Unit, $100,000.

TABLE V

THE DECILES FOR THE DATA OF TABLES 102 (1) AND 103 (2)*

Decile number	(1)	(2)
1	83	106
2	91	123
3	93	133
4	95	140
5	98	148
6	100	157
7	101	168
8	105	178
9	113	211

* Unit: per cent. Compiled from Tables 102 and 103.

TABLE W

MONTHLY AVERAGE TAX-PAID WITHDRAWALS OF LARGE CIGARS, ANNUALLY, 1919–1937*

Year	Tax-paid withdrawals
1919	589 4
1920	661 4
1921	563 2
1922	574 4
1923	583 2
1924	554.9
1925	541 7
1926	549 1
1927	547 6
1928	537 8
1929	546 0
1930	490 8
1931	443 2
1932	370 2
1933	362 1
1934	383 1
1935	397 0
1936	431 9
1937	443.1

* Unit· million cigars. Source: Annual reports for calendar years of U. S. Treasury, Bureau of Internal Revenue.

TABLE X

PETROLEUM PRODUCED IN THE UNITED STATES, ANNUALLY, 1900–1937*

Year	Production	Year	Production
1900	63 6	1919	378.4
1901	69 4	1920	442 9
1902	88 8	1921	472.2
1903	100 5	1922	557.5
1904	117.1	1923	732.4
1905	134.7	1924	713.9
1906	126 5	1925	763.7
1907	166 1	1926	770.9
1908	178.5	1927	901.1
1909	183.2	1928	901.5
1910	209.6	1929	1007.3
1911	220 4	1930	898 0
1912	222 9	1931	851.1
1913	248 4	1932	785 2
1914	265.8	1933	905.7
1915	281.1	1934	908.1
1916	300 8	1935	996.6
1917	335 3	1936	1099.7
1918	355.9	1937	1277.7

* Unit: million barrels of 42 gallons Source: "Minerals Yearbook, 1937," Washington, U. S. Department of the Interior, p. 1009, and *Monthly Petroleum Statement No.* 167, U. S. Department of the Interior, February 9, 1938.

TABLE Y

LINK RELATIVES FOR THE RATE OF INTEREST ON CALL LOANS IN THE NEW YORK STOCK EXCHANGE, 1903–1914*

Month	1903	1904	1905	1906	1907	1908	1909	1910	1911	1912	1913	1914
January	84	43	72	52	44	33	62	94	94	60	50	51
February	50	77	97	54	71	38	124	59	72	94	103	75
March	208	97	146	105	146	102	82	104	100	106	127	107
April	70	79	102	195	37	93	105	114	101	124	82	96
May	58	112	74	44	98	97	95	111	100	92	80	97
June	125	73	103	78	135	92	102	76	104	100	82	103
July	82	91	92	91	145	80	110	87	98	105	100	144
August	81	87	89	149	67	87	105	64	98	99	100	236
September	114	170	174	211	131	127	124	129	99	93	129	96
October	116	133	149	55	525	107	160	156	102	203	127	100
November	193	138	145	146	58	122	108	103	117	120	102	90
December	106	112	214	187	119	166	108	105	148	102	123	62

* Unit: 1 per cent. Source: *Review of Economic Statistics*, Prel. Vol. I, 1919, p. 103.

TABLE Z
PRODUCTION OF PORTLAND CEMENT, ANNUALLY, 1910–1926*

Year	Production
1910	76.55
1911	78 53
1912	82 44
1913	92.10
1914	88 23
1915	85 92
1916	91 52
1917	92 81
1918	71 08
1919	80 78
1920	100 02
1921	98 84
1922	114 79
1923	137 46
1924	149 36
1925	161 66
1926	164 53

* Unit· million barrels. Source: "Statistical Abstract of the United States, 1937," Washington, U. S. Department of Commerce, 1938, p 734.

TABLE AA

DAILY AVERAGE PRODUCTION OF PIG IRON, MONTHLY, 1903–1914*

Month	1903	1904	1905	1906	1907	1908	1909	1910	1911	1912	1913	1914
January	47 509	29 795	57 479	66 739	71 149	33 718	57 986	84 149	56 752	66 384	90 172	60 808
February	49 665	41 668	57 048	68 001	73 039	37 163	60 976	85 616	64 090	72.442	92 369	67 453
March	51 305	46 820	62 460	69 859	71 871	39 619	59 232	84 450	70 039	77.591	89 147	75.739
April	53 614	52 039	64 068	69 107	73 975	38 320	57 963	82.792	68 839	79 181	91.759	75.665
May	55 278	49 580	63 346	67 701	74 049	37 603	60 753	77 103	61 079	81 051	91 039	67 506
June	55 774	43 191	59 773	65 891	74 486	36 404	64 362	75 516	59 586	81 358	87..619	63.926
July	49 877	36 155	56 191	64 948	72 763	39 294	67 853	69.305	57.839	77.771	82 601	63 150
August	50 681	37 830	59 473	62 153	72 594	43 865	72 645	67 963	62 150	81 046	82 121	64 363
September	51 791	45 261	63 317	65 699	72 783	47 300	79 507	68 539	65 903	82 128	83 531	62.753
October	45 989	46 944	66 231	70 865	75 386	50 555	83 856	67 520	67 811	86 772	82 139	57 361
November	34 654	49 554	67 121	72 922	60 939	52 595	84 917	63 659	66 648	87 695	74 453	50 611
December	27 313	52 129	65 991	72 107	39 815	56 158	85 022	57.349	65 912	89 766	63 987	48 896

* Unit: thousand long tons. Source: *Iron Age*, January 3, 1935, p. 276.

TABLE BB

DAILY AVERAGE PRODUCTION OF PIG IRON, MONTHLY, 1919–1924*

Month	1919	1920	1921	1922	1923	1924
January	106.52	97.26	77.94	53.06	104.18	97.38
February	105 01	102.72	69.19	58 21	106.94	106.03
March	99 68	108 90	51.47	65.68	113 67	111.81
April	82 61	91.33	39.77	69 07	118.32	107.78
May	68 00	96 31	39.39	74 41	124.76	84 36
June	70 50	101.45	35.49	78.70	122.55	67.54
July	78 34	98 93	27 89	77.59	118.66	57.58
August	88 50	101.53	30 78	58.59	111.27	60.88
September	82 93	104.31	32 85	67 79	104.18	68.44
October	60.12	106 21	40 22	85.09	101.59	79.91
November	79 74	97.83	47.18	94.99	96 48	83.66
December	84 94	87.22	53.20	99.58	94.22	95.54

* Unit: thousand long tons. Source: *Iron Age*, January 3, 1935, p. 276.

TABLE CC

DAILY AVERAGE PRODUCTION OF BOOTS, SHOES, AND SLIPPERS, MONTHLY*

Month	1922	1923	1924	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938
January	1 092	1.282	1 104	1 109	1.061	1.086	1 140	1 135	1.105	0 846	0 944	0 988	1 085	1.232	1.390	1.651	1.110
February	1 116	1 378	1.167	1 202	1 168	1 241	1 288	1 260	1.117	1 089	1.129	1.199	1 389	1.405	1.469	1.800	1.353
March	1 174	1 463	1 228	1 245	1 198	1 251	1 318	1.315	1 218	1 223	1 227	1 143	1 451	1 457	1.451	1 845	
April	1.193	1.386	1 167	1 228	1 110	1.208	1.158	1.224	1.208	1.246	1 104	1.228	1.497	1.440	1.392	1.679	
May	1 093	1.289	1 074	1 092	1 028	1 114	1 101	1.215	1.043	1.237	0 978	1 374	1 419	1 302	1.316	1.574	
June	1.035	1.203	0.977	0 977	1 043	1.146	1 161	1.250	1 039	1.160	0 981	1.452	1 214	1 210	1.224	1.435	
July	1.008	1 098	0.892	1 011	1 066	1 234	1 223	1.259	1 005	1 168	0 908	1 500	1 234	1.345	1.456	1.483	
August	1.107	1 202	1.083	1.212	1.235	1.402	1.399	1 488	1.210	1 424	1.231	1.481	1 425	1.520	1.731	1.611	
September	1.257	1 282	1.206	1 294	1 377	1.475	1 441	1.583	1 275	1 361	1.473	1.389	1.311	1.543	1.782	1.480	
October	1 266	1 228	1 233	1.267	1 348	1 373	1 336	1.488	1 109	1 036	1.407	1.310	1 258	1.438	1.629	1 238	
November	1 308	1 172	1.178	1 120	1 163	1.129	1 150	1 232	0.862	0 842	1.093	1.030	1 037	1.232	1.380	0 926	
December	1 238	1 007	1.025	1.017	1 038	1.001	0 974	0.977	0 731	0 815	0 856	0 893	1 031	1.259	1.391	0.859	

* Unit: million pairs. Daily average production has been computed from monthly totals by dividing each monthly total by the number of working days in each month. The number of working days in each month has been obtained by making the following allowances for Sundays, Saturdays, and holidays: Sunday, 1 day; Saturday, ½ day; January 1, 1 day; May 30, 1 day; July 4, 1 day; Labor Day, 1 day; Thanksgiving, 1 day; December 25, 1 day. (If holiday falls on Sunday, observance is assumed to come on the following Monday; if holiday falls on Saturday, the holiday allowance plus the Saturday allowance is taken as 1 day.) Source: Monthly reports of the U. S. Department of Commerce, Bureau of the Census. (Rubber-soled footwear is excluded.)

TABLE DD

CYCLES OF STOCK PRICES*

Month	1919	1920	1921	1922	1923	1924
January	0 38	1.44	−0 69	−0 68	−0 12	−0 47
February	0 38	0 78	−0 70	−0 53	0 05	−0 43
March	0.70	1 06	−0.72	−0 38	0 15	−0 58
April	0.90	1 10	−0 68	−0 18	−0 02	−0 81
May	1 40	0 50	−0 67	−0 11	−0 31	−0 88
June	1 76	0 46	−1 15	−0 14	−0 46	−0 82
July	2 01	0 38	−1 20	−0 08	−0 70	−0 58
August	1.50	0 04	−1 32	0 05	−0 65	−0 39
September	1 78	0 11	−1.16	0 10	−0 70	−0 44
October	2 14	−0 04	−1 12	0 10	−0 84	−0 50
November	1 90	−0 44	−0 89	−0 16	−0 72	−0 25
December	1 54	−0 85	−0 70	−0 10	−0 60	−0 02

* Unit· one standard deviation. Source. *Review of Economic Statistics*, May 15, 1932, p. 82.

TABLE EE

HARVARD INDEX OF GENERAL ECONOMIC CONDITIONS*

Month	1925			1926			1927		
	A	B	C	A	B	C	A	B	C
January	0 00	0 47	−1 15	1 05	0 99	−0 20	1 05	0 84	−0 40
February	0 15	0 55	−1 13	1 11	0 88	−0 35	1 04	0 80	−0 59
March	0 17	0 67	−0 84	0 90	0 78	−0 39	1 31	0 79	−0 66
April	−0 21	0 46	−0 90	0 33	0 73	−0 68	1 25	0 95	−0 64
May	−0 12	0 52	−0 95	0 52	0 78	−0 70	1 32	0 88	−0 49
June	0 19	0 66	−0 89	0 55	0 81	−0 72	1 56	0 88	−0 42
July	0 27	0 70	−0 85	0 81	1 04	−0 61	1 39	0 95	−0 46
August	0 41	0 78	−0 69	1 00	0 83	−0 39	1 94	0 90	−0 78
September	0 43	0 81	−0 52	1 11	0 68	−0 26	2 07	0 99	−0 92
October	0 53	0 78	−0 30	1 13	0 70	−0 20	2 42	0 98	−0 90
November	1 01	0 96	−0 24	0 89	0 45	−0 33	2 06	0 81	−0 88
December	0 90	0.80	−0 18	1 07	0 53	−0 33	2 55	0 83	−0 89

* Unit: one standard deviation. Source. files of *Review of Economic Statistics*, by permission. (A, speculation; B, business; C, money.)

TABLE EE (*Continued*)

Month	1928			1929			1930		
	A	B	C	A	B	C	A	B	C
January	2 67	0 99	−0 67	4 41	1 52	1 89	0 74	0 67	+0.06
February	2.51	0 86	−0 58	4.21	1 61	1 89	1 11	0 62	−0 11
March	2.31	0 95	−0 58	4.11	1.73	2 04	1.28	0 67	−0.73
April	2 96	1 28	−0 29	3.70	1.59	2.49	1 83	0.67	−0 90
May	3.37	1 38	0 10	3.84	1.52	2.77	1.59	0 68	−1.11
June	3 38	1 45	0 61	3.17	1.70	2 63	1 40	0 79	−1 47
July	2.82	1 20	0 80	3 88	2.31	2 39	0.13	0 15	−1 81
August	2 89	0 98	1 02	4 25	2.13	2.73	0 25	−0 05	−2.00
September	3 37	1 28	1 23	4 89	2 13	2 70	0 24	−0 22	−2 09
October	3 63	1 21	1 30	4 10	1.96	2 04	−0 51	−0 50	−2 18
November	3 69	1 32	1 30	1 67	1.71	0 69	−1 04	−0 68	−2 37
December	4 39	1.54	1 64	0.68	0.75	0.07	−1.21	−0 72	−2 26

TABLE EE (*Continued*)

Month	1931			1932			1933		
	A	B	C	A	B	C	A	B	C
January	−1 65	−0 95	−2 46	−3 99	−2 67	−1 08	−4 37	−3 98	−4 08
February	−1 35	−1 15	−2 65	−4 02	−3 04	−1 17	−4.31	−4 11	−3.95
March	−0 83	−1 22	−2 64	−3 90	−3 42	−1 58	−4 65		−1 96
April	−1 22	−1 12	−2 75	−4 20	−3 08	−2 04	−4 62	−4 29	−3 14
May	−1 73	−1 14	−2 96	−4 64	−3 38	−2 59	−3 91	−4 08	−3.50
June	−2 35	−1 28	−3 22	−5 05	−3 55	−2 84	−3 33	−3 68	−3.67
July	−1 85	−1 60	−3 20	−5 09	−3 58	−3 06	−2 90	−3 28	−3.61
August	−2 15	−1 85	−3 23	−4 60	−3 80	−3 26	−3 26	−3 72	−3 67
September	−2 18	−1 98	−3 12	−3 86	−3 93	−3 37	−2 89	−3 89	−4 06
October	−3 42	−2 29	−1 87	−3.97	−4 04	−3 71	−3 31	−3 95	−4.07
November	−3 23	−2 68	−1.26	−4 30	−4 21	−3 92	−3 57	−3 94	−3 99
December	−3 54	−2 43	−1 34	−4 42	−3 98	−4 05	−3 33	−3 91	−3.82

TABLE EE (*Continued*)

Month	1934			1935			1936		
	A	B	C	A	B	C	A	B	C
January	−3 27	−3 83	−3 81	−3 21	−3.40	−4.13	−1.93	−2.74	−4.13
February	−2 83	−3.53	−3 88	−3.30	−3.13	−4 14	−1.61	−2.75	−4.14
March	−2.90	−3.56	−4 03	−3.37	−3.02	−4.16	−1.51	−2.65	−4.16
April	−2 89	−3.33	−4.03	−3.51	−2 93	−4.34	−1.51	−2.52	−4.16
May	−2.93	−3.33	−4.02	−3.23	−2.80	−4.51	−1.92	−2.38	−4.04
June	−3 19	−3.21	−4.13	−3.14	−2.75	−4.53	−1.71	−2.11	−4.00
July	−3 13	−3.23	−4.14	−2.97	−2.70	−4 53	−1.61	−2 02	−4.00
August	−3.51	−3.34	−4.15	−2.71	−2.72	−4.55	−1.31	−2.07	−4.02
September	−3.37	−3.38	−4.17	−2.62	−2.80	−4.55	−1.27	−2.22	−4.04
October	−3 40	−3 52	−4.17	−2.55	−2.95	−4.52	−1.23	−2.23	−4.04
November	−3 45	−3 40	−4 16	−2 29	−2 71	−4.16	−0.90	−1.96	−4.03
December	−3 21	−3.32	−4.16	−2.10	−2.60	−4.16	−0.78	−1.42	−4.03

TABLE EE (*Continued*)

Month	1937			1938		
	A	B	C	A	B	C
January	−0 81	−1 92	−4.00	−3.03	−2.53	−3.80
February	−0 64	−1.82	−4.01	−3.03	−2.78	−3.81
March	−0.61	−1.70	−3 99	−2.85	−2.72	−3.83
April	−0 65	−1.67	−3.84	−3.72		
May	−1.11	−1.54	−3.79			
June	−1.17	−1.55	−3.79			
July	−1.44	−1.55	−3 80			
August	−1 03	−1.54	−3 82			
September	−1.31	−1.77	−3.85			
October	−2 04	−1 93	−3 85			
November	−2 48	−2 33	−3.84			
December	−2 85	−2.01	−3.83			

APPENDIX B

LABORATORY PROCEDURE[1]

There are certain general desiderata for all laboratory work in statistics, and chief of these are: uniformity and adequacy of plan, neatness and accuracy in the operation, accessibility and adaptability in the results. The routine worker who fails to govern his operations by a complete plan of procedure, which conforms to some uniform standard for all operations of similar type, runs a serious risk of wasting valuable time and energy and of obtaining results which are of doubtful reliability. The final test of all detailed statistical work is accuracy—not necessarily perfect, but sufficiently close to accuracy for the purpose in hand. Careful planning of the work facilitates accurate operations, and a habit of neatness serves effectively to reduce the number of errors. The only satisfactory general safeguard against error is found in the duplicate performance of the same operation, independently, by two individuals. Comparison of the results, if it shows verification, establishes a very high probability of accuracy. The results of laboratory operations should be in such form that they can readily be found and utilized for interpretation or further analysis. The accomplishing of this end implies a foresight in planning and a thoroughness in operation which are developed only by experience but should be sought from the first.

TRANSCRIPTION

Certain general specifications can be applied to the transcription of series to be used in a statistical study. Of the three main types of series—categorical, time, and frequency—most of the published data upon which computations will be based for the analysis of an economic subject will fall within the class of time series. Undoubtedly the student will at times need to transcribe categorical series, especially if he is interested in conditions of industries or trade at a given time. However, there will be little opportunity actually to transcribe frequency series, as few series

[1] This Appendix was prepared largely, for the original edition, by Ester I. Fismer and Amelia E. Ohse, formerly of the staff of the Harvard University Committee on Economic Research.

of published economic data fall within that class, but rather need to be compiled by selection from the categorical and time series at hand. As the greatest amount of necessary data for a statistical problem falls within the class of time series, methods for transcribing these series will be described in some detail.

Transcriptions—written copies—of the figures to be used as original items in a statistical problem are almost essential. Cases arise when it seems convenient to use material directly from the source at hand without transcribing, but usually this practice is undesirable. It is most important that the original data upon which the computations, tables, and charts of a study are based be preserved; and for this reason the material should always be transcribed so that it may readily be filed with the study, for easy reference. When transcribed, the data can be arranged uniformly and thus are more convenient for use in interpretation and comparison, in drawing up tables and charts, and in computing. Another important reason for transcribing original data is that when later figures appear revising the data originally used, corrections can be entered immediately upon the original copied sheets, and these then become a record of the actual changes made in the series. These changes will then need to be carried over into any corresponding computations, tables, or charts already completed in the study.

Before transcribing data the student must be certain he has secured complete representative series from unquestionable sources, using, whenever possible, a primary source. Secondary sources containing data already tabulated in regular form are usually more easily available, but a careful selection, after research, must be made in order to choose the most reliable secondary sources. A trade journal, or publication specializing in the data desired, is often a practical solution of this problem. For example, sugar data are often taken from *The Sugar Trade Journal*, iron and steel figures from *Iron Age*, and leather prices from *Hide and Leather*.

In order to plan and arrange transcription sheets in the most convenient style, the use to be made of the data should be decided before drawing up the sheets. The form chosen depends on whether the material is to be used for computations, tables, charts, or merely for comparison and observation in deriving conclusions. When planning the sheets, the student should keep in mind any possible future use of the same data either by himself or by someone else, not only in connection with his present study, but for

other and various purposes. Therefore, the transcriber should arrange the sheets so they may easily be understood by others and so they may readily be filed in a general file. For instance, when transcribing time series of monthly production of pig iron and steel for a given number of years, it is best to draw up a separate sheet for pig iron and another for steel, so that later one series may be filed under pig iron and the other under steel.

In the case of monthly time series for a given period, such as January, 1919, to December, 1937, it is a good practice to transcribe each complete series for the specified number of years on a single sheet; but, if similar weekly series are desired, only one year for each series should appear on a sheet, as in the second illustrative table below. If, however, the material is wanted for direct comparison, two or more series may be transcribed on a single sheet. In the case of monthly series, it will then be best to limit the material to one year on a sheet, showing each individual series for the twelve months; in the case of weekly series a quarter of a year on a page is a convenient arrangement, each page showing the various series to be compared during the weeks covering three months in a year. If an index of various items is to be computed, the several series entering the index may, for the sake of convenience, be transcribed on the same sheet for comparison when finally combined in the index.

On the actual process of transcribing data the same general principles apply for all types of series. To prevent discrepancies the original data should be copied by hand, in clearly legible characters. Ink should be used, to insure neatness and to aid in preservation of data. The general tendency, when first transcribing, is to crowd many items on a single page. This frequently causes error, not only in the process of transcribing, but also by producing a confused mass of data upon which to base computations, in the subsequent steps of computing. To avoid crowding, the transcriber should spread his work out, allowing all the space necessary for average-size figures, with ample room for corrections and changes of each item. A legal-size paper, 8½ by 14 inches, most conveniently meets all the necessary requirements of transcription sheets. The following arrangements can well be observed. Space sufficient for title and captions should be allowed at the top of the sheet: for title, at least two inches vertically; for captions, not less than one vertical inch. Lines should be ruled on both sides of each column, allowing about one inch horizontally for a column of items, and more for the stubs if necessary. Not more

than twenty rows of items should be transcribed on a sheet; this spacing allows for only one revision in each space, and more space should be left if possible. At least two vertical inches should be reserved at the bottom of each sheet for footnotes. If the quality of the paper used is poor and the sheet is likely to have considerable use, leave about ½-inch margin at right and left of sheet so that the transcribed figures will not become damaged.

When the data to be transcribed are to be drawn up into tables directly for presentation in a study, or for publication, the above

COTTON CONSUMED

RAW COTTON (EXCLUDING LINTERS) CONSUMED IN THE UNITED STATES*

Month	1919	1920	1921	1922	1923	1924[a]
January	556.9	591 9	366.5	526.7	610.3	578 5
February	433.3	515 7	395.1	472 3	566 8	507 9
March	433 5	575 8	438 2	519 8	624 3	483 9
April	475 9	566 9	409.2	443 5	576 5	480 0
May	487 9	541 4	440 7	495.3	620 9	413.6
June	474.3	555.2	461.9	509.2	542.0	350.3
July	510 3	525 5	410 1	458 0	462 7	346 7
August	497 3	483.6	467.1	526 4	492 5[a]	357 5
September	491 1	458 0	484 7	494 0	485 7[a]	435 2
October	556 0	401 3	494 3	533 7	543 3[a]	532 6
November	491 3	332 7	527 9	579 2	532 7[a]	492.2
December	511.7	295 3	510 9	529 3	463.8[a]	532 0

Unit: thousand bales Data are for entire month. Source· Original data compiled by U. S Department of Commerce, Bureau of Census. Transcribed data from "Cotton Production and Distribution" Bulletins, Department of Commerce. January, 1919–July, 1919, data from Bulletin No. 150, p. 49. August, 1919–July, 1923, data from Bulletin No. 153, p. 52. Later data from monthly preliminary report postcards on cotton consumed, sent by Department of Commerce, Bureau of Census.

[a] Preliminary data to be revised when new issues of *Cotton Bulletin* are available which contain revised figures beginning in August and ending in July of the following year.

suggestions for spacings may be disregarded, and smaller spaces used. Cross-section paper will be found most convenient for this purpose, as it saves the transcriber considerable labor in drawing lines and yet produces sheets which appear clear and concise, and are properly aligned.

These principles apply to categorical series as well as time series; but in the case of monthly time series only twelve or fourteen rows of items will be needed on a sheet, to provide space for each month in the year and for the yearly total and a monthly

*Note that the form is that recommended for transcription sheets but the published table appears on a smaller scale due to the size of the printed sheet: ordinarily legal size paper should be used for transcriptions.

average if desired. In the case of weekly time series, it will be necessary to carry eight columns on a sheet in order to transcribe the data conveniently. One year appears on a single page; and there are two columns for each quarter, one for the stub, containing

AVERAGE PRICE OF COTTON*

RAW COTTON—FAIRCHILD'S AVERAGE PRICE OF NEW YORK SPOT COTTON

1924							
Jan. 5	35.65	Apr. 5	29 50	July 5	30 18	Oct. 4	26.23
12	35 00	12	31.16	12	30 26	11	25 30
19	33.64	19	30 61	19	31 84	18	23.53
26	33.48	26	29.87	26	34 60	25	23 63
Feb. 2	33.81	May 3	29 98	Aug. 2	32 24	Nov. 1	23.99
Feb. 9	34 10	May 10	30 70	Aug. 9	30.88	Nov. 8	23.83
16	32.31	17	31 59	16	29.90	15	24 68
23	30.32	24	32 23	23	27 92	22	24 26
Mar. 1	29.23	31	32 67	30	26 60	29	24.12
Mar. 8	28 48	June 7	30 71	Sept. 6	25 72	Dec. 6	23 30
15	28 79	14	29 46	13	23.93	13	23 43
22	29.00	21	29 69	20	22.53	20	24.01
29	27.27	28	29 90	27	24.46	27	24 22
						Jan. 3	24.58

Unit: cents per pound. (Average of daily prices for week ending Saturday) Source: Original data compiled by *Fairchild's News Service*. Transcribed data from: *Daily News Records* in issue of Tuesday each week appearing under table "Fairchild Cotton—Average Price and Index Number." Data revised from the annual table in *Daily News Record*, January 20, 1925.

the month and date of week, and the other for the items. Fifteen rows of items are needed for transcribing weekly time series by this method, allowing for the possible fifth week in each month of the quarter.

After the transcription sheets have been planned and prepared, the actual transcribing may be started. In the space allowed at

*Note that the form is that recommended for transcription sheets but the published table appears on a smaller scale due to the size of the printed sheet: ordinarily legal size paper should be used for transcriptions.

the top of the sheet, the title of the series, giving the description of the series in general, should be written in capital letters; and directly under the general title, the full description of the series should appear. Particular series require certain additional elements in the title. For example, in a price series the market, such as Chicago, New York, Philadelphia, and the condition of sale—that is, whether the price is spot or contract, f.o.b mill, or delivered—should be stated in the title.

Units in which the data are transcribed should also be given in a note attached to the title. The determination of the unit to be used is, of course, a step in the planning which precedes actual transcription, and depends upon the use to which the data will be put. The following suggestions may prove of assistance in selecting a unit for transcriptions of various series. In most data it is, of course, not necessary to transcribe figures to as many places as given in the actual source. If the data to be transcribed are for use in computations only, it is usually necessary to carry value and quantity figures to only three or four digits, but if desired for tables of original items and intended for comparison later in the study, it is often advisable to carry some figures to more than four digits. The reason for this is that often the number of digits in the separate items of a series will vary when they are expressed in a single unit: the larger numbers must necessarily be carried out to extra digits so that the small items may not appear too insignificant. In general, the most useful units are those of thousands and millions, which omit three digits or six, according to the size of the items in a series. Units of hundreds, ten thousands, and hundred thousands are confusing, and error is often avoided by selecting units of thousands and millions and carrying the same desired number of digits by the use of a decimal point. For example, if the production of crude petroleum in the United States for a stated period is 56,617,000 barrels, the figure can be handled more conveniently as 56.62 million barrels than by trying to use a unit of ten thousand; and, likewise, 768,903 tons of cottonseed crushed in November, 1924, is better carried as 768.9 thousand tons, than as 7,689 hundred tons, which results in the same number of digits.

The transcriber, no matter how large the actual figures are, should never under any condition use commas. If figures of many digits are to be taken off, a small space allowed between thousands and millions, as 7 835 256, is a far more satisfactory practice.

Moreover, each item should be copied, and no ditto marks should be used.

It is often advisable, when transcribing price data, to transcribe the data as actually given in the source on hand, not omitting any digits, in order to show actual price changes from time to time. For example, iron bars may appear as $0.02875 per pound in one week and $0.02865 in the next. Slight changes in a price series may be important in a study of short-time movements, and for this reason enough digits to be sensitive to small price fluctuations should be retained. When stating the unit used in a series of prices, the value per physical unit should be mentioned on the transcription sheet, as in the case of iron bars, dollars per 100 pounds, or cents per pound.

In time series, the date or interval to which the item applies should also be stated. In monthly series it is necessary to know if transcribed series are items for the entire month, first of month, end of month, or the average for the month. In weekly series it must be stated whether transcribed items are aggregates for the week, or figures as of one day in each week, or an item for the entire week ending on a particular date indicated. For example, pig iron prices in the *Iron Age* are prices as of Tuesday each week, data on car loadings, issued by the American Railway Association, are totals for the entire week ending Friday, and Fairchild's raw cotton price is a weekly average of daily prices. These descriptions of time in monthly and weekly time series are particularly helpful when using data in computing and charting, and should be clearly stated in the title, or in a note to the title, where they will not be overlooked.

Footnotes are an essential part of transcription sheets and should contain all important facts concerning the data. In the footnotes should be given the source of the data, for a record is needed of the actual origin of the transcribed material; and it should be indicated whether the source is primary or secondary. If a secondary source is used, the title, date, volume, and page of the publication should be given. Also, if more than one source is used for the data on a single sheet, the columns of items or exceptional figures should be marked accordingly, so that each item may easily be located when a check on the figures is desired. Any peculiarities in the series, such as a change in description or methods of compilation, or change in number of items included in original reporting of figures, should be noted carefully, since some

adjustment will no doubt be necessary in computations based upon such data.

The actual transcribing of data will be simple and straightforward, involving no complications, if the data are to be taken from a secondary source, where all the desired information appears in one issue of the publication, since the student can see at a glance the material desired. The work will be difficult if the items are to be taken from several issues of the same publication, and even more difficult when they are to be compiled from several sources. When transcribing series from several issues of the same publication, if no overlapping figures are given, it is necessary in each case, before taking a figure, to examine the data carefully in order to be sure the series is homogeneous. For most secondary sources, however, by selecting one item at a time from several issues of the same publication, the transcriber can secure overlapping figures; and, although it is not certain even then that the data are homogeneous, it is generally safe enough to assume homogeneity if the figures check. When overlapping figures are given, the student must use the latest publication first: that is, when a series is to be selected by following through each issue of the publication, if January to December, 1924, data are desired, he should start with the latest publication containing December data and work back to January, 1924. Each issue must be examined when working backward, from the latest to the earliest, even if considerable overlap is given in each issue; for, unless the transcriber systematically examines each issue for changes, a revision of a single item appearing in a later issue may easily be overlooked.

If the data to be transcribed are fairly old—prewar, for instance—it is easy to secure the last revised data at the present time with the assurance they will not again be changed. But if the data desired are current, and the series is one originally compiled some time ago and still being carried in the same form in a publication, the data are likely to appear revised for two or more years at a time, and need to be watched continually for revisions.

The student may be anxious to bring his study up to date before the final data appear in the publication used. Frequently, series may be completed by data appearing in press releases, newspaper clippings, typed or mimeographed preliminary reports. In this form they may readily be used, but the figures taken from such sources should be properly footnoted and marked "preliminary" (*p*). Great care should be taken to revise these figures

as soon as the usual bulletin or publication from which the series is regularly taken becomes available. The student should remember, too, that weekly data may later be revised in monthly issues or annual summaries, and monthly data from monthly publications may later be revised in annual reports issued by the same source. Care should be taken to examine recent data from time to time, and in some way to note on the sheets when the final source has been used.

In transcribing weekly series it is often difficult to specify an item for a week which apparently includes data belonging to either of two months. This problem of allocation of weeks does not arise if the weekly series is one pertaining to only one day in each week, since that item without question falls within the stated month. A figure which is an aggregate or total for a week should be transcribed as belonging to the month in which it has the greater number of days of the week.

When transcriptions are finally completed they should have, if possible, an independent check by someone other than the transcriber. Each item and footnote should be verified carefully before being used for computing or for any other purpose. When error is found or change is necessary, no erasures should be made but the erroneous figure should be crossed out neatly and the correct figure entered directly above the old. Whether revisions are due to error or simply to a revision in the data appearing in a later publication, if computations and charts based upon the data have already been completed, the student should trace through each operation affected, making the necessary corrections. For this reason, replacing figures by erasure is especially dangerous, since it cannot then plainly be seen where such changes occur. Colored ink is useful in making revisions, for the corrections then stand out clearly on the sheet and cannot easily be overlooked.

The student will be able to detect certain discrepancies in his own transcribed data by making a simple audit of the figures when transcribed. After careful examination, some errors will appear quite evident. For example, an item of a series may clearly be out of proportion with the other items (for instance, pig iron, appearing with a monthly production approximating 1,900 long tons, may suddenly appear as 900 long tons). This item needs examination until the transcriber is satisfied that cause exists for such an irregular shift. The explanations for some of the peculiarities in a series are often given in the text accompanying the figures in the source, but if no real reason can be found, the cause

may often quite clearly be due to typographical error. In such cases, the data should be verified elsewhere, either by writing directly to the compilers and questioning the figures, or by referring, whenever possible, to another reliable source.

Many of the foregoing points are illustrated in the two tables shown above (pages 394 and 395).

COMPUTATION

Written instructions should form a part of every computing job: they not only facilitate the planning of the work but are valuable for future reference, because they serve as a record of what has been done. These instructions should give specific directions as to the data to be used, the precise methods to be applied, and the results desired. For example, if a series is to be corrected for trend and seasonal variation and then reduced to cycle form, the instructions should state the type of trend line to be fitted and the period on which the trend should be based; the exact period for which the link relatives are to be computed, and the period on which the standard deviation is to be based. In general, when time series are used, it is necessary that the instructions specify the period of time to which the figures pertain. In fact, any information the computer will need in connection with the computation should be stated clearly. If the instructions have been set down by someone else, it is very important that the computer study and analyze carefully all the directions before he attempts to draw up the computation sheets. If this is done, he will have the entire job clearly in mind and can save himself hours of energy wasted in preparing inappropriate sheets.

The principles involved in the drawing up of computation sheets are very similar to those mentioned in the section pertaining to the drawing up of transcription sheets; but as they are very important, they may bear some repetition. Of these principles, neatness is the most important because it tends to ensure accuracy; and if all work is not done neatly, the result will be poorly arranged sheets, which bring about confusion and cause error. All computations which are to be used a great deal should be done in ink, to facilitate preservation. As in transcribing, all figures should be written legibly. Crowding should be avoided; ample space should be allowed for possible corrections and revisions.

Another detail to which the computer must give attention is the size of computation sheets. These sheets should be of uniform size, preferably 8½ inches by 14 inches, and should allow ample

space for headings, footnotes, and wide columns for computations. Small pieces of paper are unsatisfactory, except for scratch work to be thrown away, because they naturally become crowded, are harder to handle, and are easily lost.

No general rule can be given for the inclusion of specific operations upon specific sheets in computation. Sometimes it is desirable to arrange all of one step on a single sheet, as in computing the link relatives for a given monthly series, or in deriving the averages or aggregates of each of several series. Sometimes it is preferable to arrange several steps, in the analysis of one series or one section of a series, on a single page. For example, if a series is to be corrected for trend and seasonal variation, there should be a sheet for each year. Each sheet would then have a column for original items, one for ordinates of trend, one for ratios of original items to ordinates of trend, one for seasonal indexes, one for percentage deviations, and one for the cycles (or adjusted relatives, if preferred). This method of arranging several steps on one sheet is also desirable when an average or index is computed from the items of several series. In such instances there would be a column for each series, one for the total—whether weighted or simple—of all the series, and one for the average or index. In the analysis of frequency series, it is seldom feasible to place more than one series on a sheet, but often several operations can be applied to a single series.

On the drawing up of computation sheets a final principle to observe is that each sheet used in connection with a computation job should bear a title giving a concise but complete description of the figures appearing on the sheet, a statement of the unit in a conspicuous place, and a statement of the period of time to which the series pertain in case of time series; and the sheet should carry such footnotes as are essential. Essential footnotes include such matters as: change in description; break in series; omission of data; periods on which trend, seasonal indexes, and standard deviation are based; and similar information. In fact, any footnotes which appear on the transcription sheets and affect the original items in any way should appear again on the computation sheets. Although footnotes are important, it is not necessary to enter more than once those which are obviously repeated. It is generally sufficient that one footnote referring to several years be given in detail but once, on the first sheet, or that a note affecting each of several series be stated in full but once. For any succeeding use of the same footnote, the computer may simply refer to it.

After the instructions are clearly in mind and the computation sheets drawn up, actual computing can be started. Accuracy is most essential, and in order to avoid discrepancies, every precaution should be taken. To detect errors, two independent computations should be made. This double system may be carried out by having the two computations done on different-colored sheets, one color being considered the first copy throughout the successive steps, and the other color the second copy; or, the two copies may be done on the same-colored sheets and designated as first and second copies throughout. In cases where there is only one person to do the work, two independent computations may be secured by employing different methods. For example, after one has secured annual averages by dividing the totals of 12 monthly figures by 12, an independent check would consist in multiplying these averages by 12, and then checking the products against the yearly totals.

It is not sufficient that there be two independent computations; these computations should be kept separate throughout successive steps. That is, the person who is working on the first copy should take his data for earlier steps from the first copy and not refer to the second copy, whereas the person working on the second copy should take his figures for earlier steps from the second copy and not refer to the first copy. There are, however, exceptions to this rule, arising ordinarily in cases where a particular step is quite distinct from other operations. For example, in the computation of seasonal indexes, it is the general practice to use what is considered the first copy of link relatives in computing both copies of the medians, and also to use the medians from the first copy in getting the final indexes. The usual reason for making this exception is to prevent cumulative errors.

After the two computations have been completed, they should be carefully compared with each other. The degree of error to be allowed will depend upon the mechanical devices used in computing. For example, if ratios are read to the nearest percentage on a 10-inch slide rule, a discrepancy of 1 per cent should be allowed; whereas, if the ratios are read to tenths on a 20-inch slide rule, a discrepancy of not more than $\frac{3}{10}$ of 1 per cent should be allowed. Any discrepancies falling outside the recognized limits should be noted and corrected. All corrections should be made by drawing a neat line through the incorrect figures and writing the correct figures above—never by erasing, writing over, or altering the figures in any way. In all computation work, it is a good practice to check one step before doing successive steps, in order to avoid

the necessity of carrying possible corrections through later steps already computed.

The decision as to the number of decimal places to carry in intermediate steps cannot be made by a few general rules. The degree of accuracy in the given data sets a maximum upon the accuracy of operations applied to such data. On the other hand, the degree of accuracy desired in the results serves to specify a minimum of accuracy for the steps in computing. Within these limits, the computer must make his decision; and experience will be his chief guide. Examination of the problem will often afford considerable help in deciding how precisely to handle the arithmetic. If the operation to be applied is simple and limited to one or two steps, the answer is usually easy to find. The fundamental algebraic operations have certain calculable effects upon the errors inherent in the data, and allowance can easily be made in these cases.

When several steps are comprised in the operation, it is difficult to assess the manner in which error will be transmitted. An actual test may need to be made, in order to discover how great an allowance for error will be necessary. Frequent use of a particular process of computation may lead to fairly well defined rules for the limitation of error to a specified degree. For example, in computing original items and ordinates of trend in a time series, three significant digits, when the first digit is two or more, are sufficient, and four digits when the first digit is one. For relatives, percentage deviations, and seasonal indexes, it is generally sufficient to carry figures to the nearest per cent only. In cases where the variations are very small, however, one decimal place would be necessary. When deriving the standard deviation of cyclical movements, the percentage deviations should be read to the nearest per cent before squaring, unless they are very small, and then it is essential to take them to tenths before squaring. It is sufficient to carry cycles to one decimal place.

There are a few miscellaneous matters, such as revising data, marking estimated figures, care to use figures from original sheets, and the grouping of overlapping weeks into months, which should be considered. In revising data, the computer should use colored ink, distinct from that used in the original computation, in order to make the revisions stand out clearly. He should also make certain that these revisions have been carried through every necessary step of the computation. If the revision affects the first step but does not alter the succeeding steps, the latter should be marked with a small (*r*) tc show that the revision has been noted.

It is also important to mark all estimated figures. This should be done in cases where the computation is based on data which are not complete for the entire period in question. For example, if a monthly total for car loadings is desired and the data for the entire month are not available at the time of computation, the figure obtained by adding an estimated figure for the missing days to that already available is only a preliminary figure and should, therefore, be marked "estimated." A simple plan is to mark all estimated figures with a small (*e*). Figures should be taken from original sheets. The computer should always refer to original transcriptions and computations rather than to copies.

Another problem which frequently arises in computing work is the method of grouping into months overlapping weekly figures which pertain to an entire week. There are two methods of allocating the overlapping week; one is to assign the entire overlapping week to either the preceding or the following month, and the other is to divide the overlapping week between the two months according to the number of its calendar days falling in each month. The first method is generally used in computing monthly averages of weekly figures, while the second method is more frequently used in securing monthly totals. If the overlapping week is to be allocated according to the first method, the practice generally followed is to assign the week in question to the month in which it has more days. In case the overlapping week has as many days in the preceding month as in the following month, it does not matter in which month it is grouped, provided the same practice is followed throughout the series. When overlapping weeks are allocated by dividing them between two months, holidays should be taken into consideration. For example, if the weekly figure for bank debits pertains to the week ending January 3, this overlapping week should be allocated by distributing three-fifths of the figure to December and two-fifths to January; that is, January 1, being a holiday, must be taken out before allocating.

As an aid in statistical work, the computer will find several mechanical devices will prove helpful. For ordinary multiplying and dividing, when one does not require greater accuracy than three digits, a 10-inch slide rule will be found sufficient. If, however, a degree of accuracy over three digits is desired, it becomes necessary to resort to a 20-inch slide rule for four-digit results, and a cylindrical rule, such as the Thacher, for five-digit results. An adding machine becomes almost essential if there is much adding or subtracting to be done, and a machine which lists is desirable.

Listing is an important function, because it makes possible the preservation of a record of the adding-machine work. Many special time-saving schemes can be worked out for such a machine. For example, in computing ordinates of trend, the computer can set down the first ordinate, clear the machine, push down the repeat key, and then set down the monthly or annual increment and by one touch of the lever secure the successive ordinate. For the greater part of computation work, a slide rule and adding machine are sufficient mechanical equipment, but if more nearly exact results in either multiplication or division are desired, one of the standard calculating machines is helpful.

CHARTING

The bulk of the charting work done in the laboratory will consist in the preparation of study charts rather than of carefully finished charts for publication. Although the same care as respects plan and accuracy is essential in a study chart as in a presentation chart, much of the labor incident to preparing for publication can appropriately be saved in preparing charts of the study type. Whether a chart is to be in pencil or in ink generally depends on whether it is for study or publication, although study charts are sometimes drawn up in ink. In any case, the plotting is done in pencil to begin with, and the procedure is therefore general as far as the completion of the pencil work.

The preparation of a particular chart should be covered by written instructions, and the draftsman should study all the directions carefully before undertaking the task. Decisions must then be made as to the size of the chart, the form of grid rulings to be used, and the scales upon which the measurements are to be made. If the grid is not one of the standard types, it must be ruled up before any plotting can be done. If the chart is to be drawn on transparent paper or cloth, a ruled background may be used as grid. This practice effects a great saving in labor. In general, ruled backgrounds or other grid layouts should be few in number and simple in form. For plotting time series, the choice of time scale is generally limited to years, quarters, months, or weeks. For annual series, vertical rulings are needed for each year, with perhaps each fifth ruling heavier than the rest. For monthly or quarterly series, the annual rulings should be drawn heavy. For weekly series, monthly and annual rulings will not fall exactly on weekly rulings, and should be carefully located. For locating points vertically, a rather closely spaced set of horizontal rulings, either on an arithmetic or on a log scale, is needed, unless each

measurement is to be made with a movable scale. In case a log scale is desired which has a 1:10 ratio in some other length than the standard 5-inch, 10-inch, or 20-inch lengths taken from slide rules, such scale can be made by looking up logarithms, or by adapting a standard scale by the parallel-line construction learned in plane geometry.

For plotting many categorical series and frequency series, particularly of the block diagram type, a background ruled in squares of rather small size is generally effective. If the rulings on such a background are simple fractions of an inch or centimeter, interpolation between the rulings is easy by use of ordinary measuring sticks. Here, as in the case of time series grids, the student will do well to have only a minimum number of different grids; and he will find most plotting can be handled without special and peculiar grid plans.

As soon as the grid is established, the individual points are plotted and other essential measurements laid off. If the chart is designed to show graphic curves, the individual points should be joined by pencil lines. If the chart is of some other type, the appropriate forms should be filled in; for example, the bars of a bar diagram should be filled in with pencil lines. Individual curves (or other forms on the chart) should be labeled with appropriate captions, the scale numbers should be entered, and the title marked upon the chart. The nature and source—the latter presumably a working table—should be noted.

It is then time for the entire chart to be verified, preferably by some one other than the draftsman. The verification consists in checking every part of the chart—plotted points, curves, ruled forms, captions, scale numbers, and title—in the light of the instructions and the original data.

The degree of error to allow depends largely upon the use to which the chart will be put. Perfect graphic presentation is impossible, but a range of allowable error should be determined in advance and adhered to for each chart. Ultimately the degree of accuracy demanded is decided upon the same grounds as the degree of accuracy used in computing work. In charting, the weight of the curve in the final chart limits somewhat the precision with which points are shown, but fundamentally the precision is determined by the inherent accuracy of the data and the purpose of the graphic presentation. As soon as the study chart is verified in pencil form, it is finished and ready for use. Neatness and a passable regularity can be achieved by practice and the exercise of constant care.

APPENDIX C

THE NORMAL CURVE OF ERROR, AND OTHER FREQUENCY CURVES

THE PARAMETER OF THE CURVE

It has been seen (page 209) that the equation of the normal curve of error has the form:

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

where N is the total frequency, σ is the standard deviation of the "errors" (deviations from the mean) x, e is the Napierian base

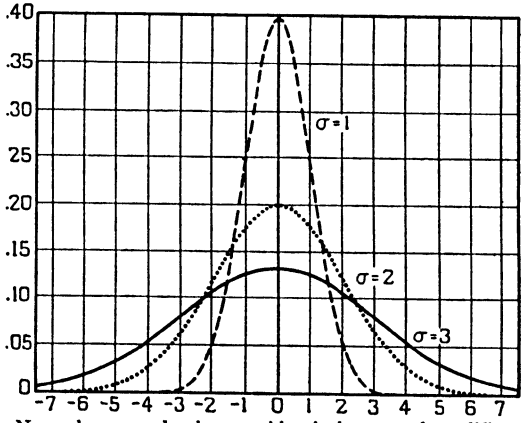


CHART 131.—Normal curves having an identical mean, but different standard deviations.

(Data in Table 147.)

2.71828, and π is the circular constant 3.14159. The equation can be written

$$\frac{y}{N} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

which gives an expression for the *relative* (per cent of the total) frequency y/N. So far as the form of the curve is concerned, the

actual frequency is of no importance, and, therefore, N is not a determinant of the form of the curve.

The forms of two normal curves each showing relative frequencies y/N (or of two curves showing actual frequencies y, if the total frequency N is the same for both curves) differ according as σ is different for the two curves. For this reason σ is called the parameter of the curve. Chart 131 (based on Table 147) shows three normal curves, all having the same N, but different values of σ.

THE QUARTILES OF THE NORMAL CURVE

If the curve is taken in the second form:

$$Y = \frac{y}{N} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

the area under the curve and between the ordinates at the quartiles must equal ½. Therefore, if Q is the quartile, then

$$\int_0^Q \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx = \frac{1}{\sigma\sqrt{2\pi}} \int_0^Q e^{-\frac{x^2}{2\sigma^2}} dx = \frac{1}{4}$$

If the variable is changed to z by means of the substitution

$$z = \frac{x}{\sigma\sqrt{2}}$$

and if

$$q = \frac{Q}{\sigma\sqrt{2}}$$

the above equation becomes

$$\frac{2}{\sqrt{\pi}} \int_0^q e^{-z^2} dz = \frac{1}{2}$$

The value of q, from Peirce's *Table of Integrals*, is found to be 0.47693. And, therefore,

$$Q = \sigma q\sqrt{2} = 0.6744^{+}\sigma$$

The values of the probability integral may be conveniently expressed as

$$\frac{1}{\sqrt{2\pi}} \int_0^u e^{-\frac{u^2}{2}} du$$

the values of which for various values of the limit u are given in Table 145.

TABLE 145

AREAS AND ORDINATES OF THE PROBABILITY CURVE*

u	$\frac{1}{\sqrt{2\pi}}\int_0^u e^{-\frac{u^2}{2}}du$ a	$\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$ b	u	$\frac{1}{\sqrt{2\pi}}\int_0^u e^{-\frac{u^2}{2}}du$ a	$\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$ b
0 00	0 0000	0 3989	2 00	0 4773	0.0540
0 04	0 0160	0 3986	2 04	0 4793	0 0498
0 08	0 0319	0 3977	2 08	0 4812	0.0459
0 12	0 0478	0 3961	2 12	0.4830	0 0422
0.16	0 0636	0 3939	2.16	0 4846	0.0387
0 20	0 0793	0.3910	2 20	0 4861	0 0355
0 24	0 0948	0 3876	2 24	0 4875	0 0325
0.28	0 1103	0 3836	2 28	0 4887	0 0297
0.32	0 1255	0 3790	2 32	0.4898	0 0271
0.36	0 1406	0 3739	2.36	0.4909	0.0246
0 40	0 1554	0 3683	2.40	0.4918	0 0224
0.44	0 1700	0 3621	2 44	0 4927	0 0203
0 48	0 1844	0 3555	2 48	0 4934	0.0184
0 52	0 1985	0 3485	2 52	0 4941	0.0167
0 56	0 2123	0.3411	2.56	0.4948	0.0151
0 60	0 2258	0 3332	2 60	0 4953	0 0136
0 64	0 2389	0 3251	2 64	0 4959	0 0122
0 68	0 2518	0 3166	2 68	0 4963	0 0110
0.72	0 2642	0 3079	2 72	0.4967	0 0099
0.76	0 2764	0 2989	2 76	0 4971	0 0089
0 80	0 2881	0 2897	2 80	0 4974	0 0079
0 84	0 2996	0 2803	2 84	0 4977	0 0071
0.88	0 3106	0.2709	2 88	0 4980	0 0063
0 92	0 3212	0 2613	2 92	0 4983	0.0056
0 96	0 3315	0.2516	2 96	0 4985	0.0050
1 00	0 3413	0 2420	3 00	0 4987	0 0044
1 04	0 3508	0 2323	3 04	0.4988	0 0039
1 08	0 3599	0 2227	3.08	0.4990	0 0035
1 12	0 3686	0 2131	3 12	0 4991	0 0031
1 16	0 3770	0 2036	3.16	0 4992	0 0027
1 20	0 3849	0 1942	3 20	0 4993	0.0024
1.24	0 3925	0 1849	3 24	0 4994	0 0021
1 28	0 3997	0 1759	3 28	0 4995	0 0018
1 32	0 4066	0.1669	3 32	0 4996	0.0016
1.36	0 4131	0.1582	3.36	0 4996	0.0014
1 40	0 4192	0.1497	3 40	0 4997	0.0012
1.44	0 4251	0.1415	3.44	0 4997	0.0011
1 48	0 4306	0.1334	3 48	0 4998	0.0009
1.52	0 4357	0 1257	3.52	0 4998	0 0008
1 56	0 4406	0.1182	3.56	0 4998	0 0007
1 60	0 4452	0 1109	3 60	0 4998	0.0006
1 64	0 4495	0 1040	3 64	0 4999	0 0005
1 68	0 4535	0 0973	3 68	0 4999	0 0005
1 72	0 4573	0 0909	3 72	0 4999	0.0004
1 76	0 4608	0.0848	3.76	0 4999	0.0003
1 80	0 4641	0 0790	3 80	0.4999	0.0003
1 84	0 4671	0.0734	3 84	0 4999	0 0003
1 88	0 4700	0 0681	3 88	0.5000	0 0002
1 92	0 4726	0 0632	3.92	0 5000	0 0002
1.96	0 4750	0 0584	3.96	0 5000	0 0002

* Adapted by permission from Professor James W. Glover's *Tables of Applied Mathematics, Finance, Insurance, Statistics;* published at Ann Arbor, George Wahr, in 1923.

THE AVERAGE DEVIATION FOR THE NORMAL CURVE

The average deviation, *A.D.*, is obtained by summing all the deviations, without regard to sign, and dividing by the total frequency. Hence

$$A.D. = 2\int_0^\infty \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} x\, dx$$
$$= \sigma\sqrt{\frac{2}{\pi}}\int_0^\infty e^{-\frac{z^2}{2}} z\, dz, \text{ in which } z = \frac{x}{\sigma}$$
$$= \sigma\sqrt{\frac{2}{\pi}}$$
$$= .7979\sigma$$

Hence, almost exactly, the average deviation of the normal distribution is 0.8 of the standard deviation.

DISTRIBUTIONS WHICH ARE NEARLY NORMAL

The relations between the standard deviation and the quartile deviation, and the average deviation

$$Q = .6745\sigma$$
$$A.D. = .7979\sigma$$

are strictly applicable only to normal distributions. In general, also, distributions which are not normal do not have mean, mode, and median coincident, as is the case with normal distributions.

Ordinarily, the nearest approach to a normal distribution realized in actual statistics appears in the case of data arising in games of pure chance. For example, the data of Table 146 form several series each of which is nearly normal. Even in these cases considerable departures from the error curve appear, but such departures are largely, if not entirely, ascribable to the imperfections of the sampling process. In other words, it may be assumed that, if any one of the series of Table 146 were actually a perfect sample of the underlying frequency series representing the true chance of turning heads, such series would be normal.

A series of data arising in the process of "observation" of a particular magnitude in physical or natural science is also likely to be so nearly normal that the properties of the normal curve may be assumed applicable. It is on this ground that the more common methods of adjustment of the errors of observation have developed about the notion of least squares.

In economic statistics there are numerous instances in which frequency distributions are very nearly normal, but there are great numbers of economic series for which the departure from normal is so great that the properties of the normal curve are not even roughly applicable.

It has been stated in the text (Chap. XIII) that even for a large number of these asymmetrical series the theory of sampling is approximately valid, especially for the simpler characteristics

TABLE 146

DISTRIBUTION OF THE NUMBER OF HEADS RESULTING FROM 100 TOSSES OF 24 COINS, FOR EACH OF 10 SEPARATE SITTINGS

Number of heads	Sitting										Total
	A	B	C	D	E	F	G	H	I	J	
0	...	...	...	...	...	...	...	. .	..	...	.....
1	...	...	...	...	...	...	...	...	...	. .	.....
2	...	...	...	...	...	...	...	...	...	...	.....
3	...	...	...	...	...	...	...	...	...	.	.....
4	...	...	...	...	...	...	...	...	...	..	.....
5	1	...	...	...	...	2	. .	.	..	...	3
6	...	2	1	...	...	2	2	3	. .	1	11
7	2	...	1	...	3	3	1	2	4	..	16
8	4	4	3	6	6	5	4	4	3	4	43
9	8	7	10	5	7	7	14	6	5	10	79
10	9	10	16	13	14	8	11	11	15	12	119
11	11	17	17	14	11	14	14	19	14	15	146
12	25	18	16	19	14	14	16	12	18	14	166
13	17	18	12	20	15	16	16	10	20	18	162
14	9	13	11	7	12	11	12	17	9	13	114
15	9	5	5	6	8	6	6	10	6	6	67
16	. .	2	2	6	5	4	2	5	4	5	35
17	3	2	5	2	3	4	2	1	2	2	26
18	1	2	1	1	...	2	..		.	.	7
19	1	...	...	...	1	2	. .	.			4
20	..	...	...	1	1	.	...	.	.	.	2
21	..	.	.	..	. .	.	.	..	.	.	.....
22	. .	...	...	.	.	.	..	...		..	.....
23	..	. .	..	..		..		.	.	.	.....
24	. .	. .	.	.	...	...	. .	. .		.	.....
Total	100	100	100	100	100	100	100	100	100	100	1,000

such as the mean. The probable errors of the principal frequency constants, including certain of the characteristics of correlation distributions, have been worked out. Although these probable error formulas have been derived for normal distributions, they can be applied, with appropriate reservations, to series which differ considerably from normal.

The application, to series customarily encountered in economics, of more refined curve-fitting methods than the mere determination of the normal curve which best fits the given data

is not ordinarily justified. Many economic series are so extremely skew that good fits are not readily obtained by any method, and interpretation of the resulting mathematical formulas is in all cases more difficult than for the case of the normal curve.

TABLE 147

ORDINATES OF THE NORMAL CURVE FOR THREE DISTINCT VALUES OF THE STANDARD DEVIATION

Total area under each curve is unity. See Chart 131

x	$y = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$		
	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$
0 00	0 399	0 199	0 133
0 15	0 394	0 199	0 133
0 30	0 381	0 197	0 132
0 45	0 361	0 194	0 131
0 60	0 333	0 191	0 130
0 75	0 301	0 186	0.129
0 90	0 266	0 180	0 127
1 05	0 230	0 174	0 125
1.20	0 194	0 167	0 123
1 35	0 160	0 159	0.120
1 50	0 130	0 151	0 117
1 65	0 102	0 142	0 114
1 80	0 079	0 133	0 111
1.95	0 060	0 124	0 108
2 10	0 044	0 115	0 104
2 25	0 032	0 106	0 100
2 40	0 022	0 097	0 097
2 55	0 015	0 088	0 093
2 70	0.010	0 080	0 089
2 85	0 007	0 072	0 085
3 00	0 004	0 065	0 081
3 15	0 003	0 058	0 077
3 30	0 002	0 051	0 073
3 45	0 001	0 045	0 069
3.60	0.001	0 039	0.065
3.75	0.000	0.034	0.061
3 90	. ..	0 030	0 057
4 05	.....	0 026	0.053
4 20	...	0 022	0 050
4.35	.....	0.019	0.046

APPENDIX D

LOGARITHMS OF NUMBERS*

Pages 413 and 414 give the common logarithms of numbers between 1 and 10, correct to four places. Moving the decimal point n places to the right (or left) in the number is equivalent to adding n (or $-n$) to the logarithm. Thus, log 0.017453 = 0.2419 − 2 (= $\bar{2}$.2419).

	0	1	2	3	4	5	6	7	8	9	10	Tenths of the Tabular Difference 1 2 3 4 5
1.0	0 0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	0414	
1.1	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	0792	
1.2	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	1139	
1.3	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	1461	To avoid interpolation in the first ten lines, use the special table on pages 415 and 416.
1.4	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	1761	
1.5	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	2041	
1.6	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	2304	
1.7	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2553	
1.8	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2788	
1.9	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	3010	
2.0	0 3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	3222	2 4 6 8 11
2.1	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	3424	2 4 6 8 10
2.2	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	3617	2 4 6 8 10
2.3	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	3802	2 4 5 7 9
2.4	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	3979	2 4 5 7 9
2.5	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	4150	2 3 5 7 9
2.6	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	4314	2 3 5 7 8
2.7	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	4472	2 3 5 6 8
2.8	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	4624	2 3 5 6 8
2.9	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	4771	1 3 4 6 7
3.0	0 4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	4914	1 3 4 6 7
3.1	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	5051	1 3 4 6 7
3.2	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	5185	1 3 4 5 7
3.3	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	5315	1 3 4 5 6
3.4	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	5441	1 3 4 5 6
3.5	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	5563	1 2 4 5 6
3.6	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	5682	1 2 4 5 6
3.7	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	5798	1 2 3 5 6
3.8	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	5911	1 2 3 5 6
3.9	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	6021	1 2 3 4 6
4.0	0.6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	6128	1 2 3 4 5
4.1	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	6232	1 2 3 4 5
4.2	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	6335	1 2 3 4 5
4.3	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	6435	1 2 3 4 5
4.4	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522	6532	1 2 3 4 5
4.5	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	6628	1 2 3 4 5
4.6	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	6721	1 2 3 4 5
4.7	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	6812	1 2 3 4 5
4.8	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	6902	1 2 3 4 4
4.9	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	6990	1 2 3 4 4

*Taken, with permission, from E. V. Huntington's *Four Place Tables*, published by Houghton Mifflin Company.

	0	1	2	3	4	5	6	7	8	9	10	Tenths of the Tabular Difference 1	2	3	4	5
5.0	0.6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	7076	1	2	3	3	4
5.1	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	7160	1	2	3	3	4
5.2	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	7243	1	2	2	3	4
5.3	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	7324	1	2	2	3	4
5.4	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	7404	1	2	2	3	4
5.5	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	7482	1	2	2	3	4
5.6	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	7559	1	2	2	3	4
5.7	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	7634	1	2	2	3	4
5.8	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	7709	1	1	2	3	4
5.9	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	7782	1	1	2	3	4
6.0	0.7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	7853	1	1	2	3	4
6.1	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	7924	1	1	2	3	4
6.2	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	7993	1	1	2	3	3
6.3	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	8062	1	1	2	3	3
6.4	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	8129	1	1	2	3	3
6.5	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	8195	1	1	2	3	3
6.6	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	8261	1	1	2	3	3
6.7	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	8325	1	1	2	3	3
6.8	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	8388	1	1	2	3	3
6.9	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	8451	1	1	2	3	3
7.0	0.8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	8513	1	1	2	2	3
7.1	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	8573	1	1	2	2	3
7.2	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	8633	1	1	2	2	3
7.3	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	8692	1	1	2	2	3
7.4	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	8751	1	1	2	2	3
7.5	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	8808	1	1	2	2	3
7.6	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	8865	1	1	2	2	3
7.7	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	8921	1	1	2	2	3
7.8	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	8976	1	1	2	2	3
7.9	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	9031	1	1	2	2	3
8.0	0.9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	9085	1	1	2	2	3
8.1	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	9138	1	1	2	2	3
8.2	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	9191	1	1	2	2	3
8.3	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	9243	1	1	2	2	3
8.4	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	9294	1	1	2	2	3
8.5	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	9345	1	1	2	2	3
8.6	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	9395	1	1	2	2	3
8.7	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	9445	0	1	1	2	2
8.8	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	9494	0	1	1	2	2
8.9	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	9542	0	1	1	2	2
9.0	0.9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	9590	0	1	1	2	2
9.1	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	9638	0	1	1	2	2
9.2	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	9685	0	1	1	2	2
9.3	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	9731	0	1	1	2	2
9.4	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	9777	0	1	1	2	2
9.5	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	9823	0	1	1	2	2
9.6	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	9868	0	1	1	2	2
9.7	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	9912	0	1	1	2	2
9.8	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	9956	0	1	1	2	2
9.9	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996		0	1	1	2	2

	0	1	2	3	4	5	6	7	8	9	10
1.00	0.0000	0004	0009	0013	0017	0022	0026	0030	0035	0039	0043
1.01	0043	0048	0052	0056	0060	0065	0069	0073	0077	0082	0086
1.02	0086	0090	0095	0099	0103	0107	0111	0116	0120	0124	0128
1.03	0128	0133	0137	0141	0145	0149	0154	0158	0162	0166	0170
1.04	0170	0175	0179	0183	0187	0191	0195	0199	0204	0208	0212
1.05	0212	0216	0220	0224	0228	0233	0237	0241	0245	0249	0253
1.06	0253	0257	0261	0265	0269	0273	0278	0282	0286	0290	0294
1.07	0294	0298	0302	0306	0310	0314	0318	0322	0326	0330	0334
1.08	0334	0338	0342	0346	0350	0354	0358	0362	0366	0370	0374
1.09	0374	0378	0382	0386	0390	0394	0398	0402	0406	0410	0414
1.10	0414	0418	0422	0426	0430	0434	0438	0441	0445	0449	0453
1.11	0453	0457	0461	0465	0469	0473	0477	0481	0484	0488	0492
1.12	0492	0496	0500	0504	0508	0512	0515	0519	0523	0527	0531
1.13	0531	0535	0538	0542	0546	0550	0554	0558	0561	0565	0569
1.14	0569	0573	0577	0580	0584	0588	0592	0596	0599	0603	0607
1.15	0607	0611	0615	0618	0622	0626	0630	0633	0637	0641	0645
1.16	0645	0648	0652	0656	0660	0663	0667	0671	0674	0678	0682
1.17	0682	0686	0689	0693	0697	0700	0704	0708	0711	0715	0719
1.18	0719	0722	0726	0730	0734	0737	0741	0745	0748	0752	0755
1.19	0755	0759	0763	0766	0770	0774	0777	0781	0785	0788	0792
1.20	0.0792	0795	0799	0803	0806	0810	0813	0817	0821	0824	0828
1.21	0828	0831	0835	0839	0842	0846	0849	0853	0856	0860	0864
1.22	0864	0867	0871	0874	0878	0881	0885	0888	0892	0896	0899
1.23	0899	0903	0906	0910	0913	0917	0920	0924	0927	0931	0934
1.24	0934	0938	0941	0945	0948	0952	0955	0959	0962	0966	0969
1.25	0969	0973	0976	0980	0983	0986	0990	0993	0997	1000	1004
1.26	1004	1007	1011	1014	1017	1021	1024	1028	1031	1035	1038
1.27	1038	1041	1045	1048	1052	1055	1059	1062	1065	1069	1072
1.28	1072	1075	1079	1082	1086	1089	1092	1096	1099	1103	1106
1.29	1106	1109	1113	1116	1119	1123	1126	1129	1133	1136	1139
1.30	0.1139	1143	1146	1149	1153	1156	1159	1163	1166	1169	1173
1.31	1173	1176	1179	1183	1186	1189	1193	1196	1199	1202	1206
1.32	1206	1209	1212	1216	1219	1222	1225	1229	1232	1235	1239
1.33	1239	1242	1245	1248	1252	1255	1258	1261	1265	1268	1271
1.34	1271	1274	1278	1281	1284	1287	1290	1294	1297	1300	1303
1.35	1303	1307	1310	1313	1316	1319	1323	1326	1329	1332	1335
1.36	1335	1339	1342	1345	1348	1351	1355	1358	1361	1364	1367
1.37	1367	1370	1374	1377	1380	1383	1386	1389	1392	1396	1399
1.38	1399	1402	1405	1408	1411	1414	1418	1421	1424	1427	1430
1.39	1430	1433	1436	1440	1443	1446	1449	1452	1455	1458	1461
1.40	0.1461	1464	1467	1471	1474	1477	1480	1483	1486	1489	1492
1.41	1492	1495	1498	1501	1504	1508	1511	1514	1517	1520	1523
1.42	1523	1526	1529	1532	1535	1538	1541	1544	1547	1550	1553
1.43	1553	1556	1559	1562	1565	1569	1572	1575	1578	1581	1584
1.44	1584	1587	1590	1593	1596	1599	1602	1605	1608	1611	1614
1.45	1614	1617	1620	1623	1626	1629	1632	1635	1638	1641	1644
1.46	1644	1647	1649	1652	1655	1658	1661	1664	1667	1670	1673
1.47	1673	1676	1679	1682	1685	1688	1691	1694	1697	1700	1703
1.48	1703	1706	1708	1711	1714	1717	1720	1723	1726	1729	1732
1.49	1732	1735	1738	1741	1744	1746	1749	1752	1755	1758	1761

	0	1	2	3	4	5	6	7	8	9	10
1.50	0.1761	1764	1767	1770	1772	1775	1778	1781	1784	1787	1790
1.51	1790	1793	1796	1798	1801	1804	1807	1810	1813	1816	1818
1.52	1818	1821	1824	1827	1830	1833	1836	1838	1841	1844	1847
1.53	1847	1850	1853	1855	1858	1861	1864	1867	1870	1872	1875
1.54	1875	1878	1881	1884	1886	1889	1892	1895	1898	1901	1903
1.55	1903	1906	1909	1912	1915	1917	1920	1923	1926	1928	1931
1.56	1931	1934	1937	1940	1942	1945	1948	1951	1953	1956	1959
1.57	1959	1962	1965	1967	1970	1973	1976	1978	1981	1984	1987
1.58	1987	1989	1992	1995	1998	2000	2003	2006	2009	2011	2014
1.59	2014	2017	2019	2022	2025	2028	2030	2033	2036	2038	2041
1.60	0.2041	2044	2047	2049	2052	2055	2057	2060	2063	2066	2068
1.61	2068	2071	2074	2076	2079	2082	2084	2087	2090	2092	2095
1.62	2095	2098	2101	2103	2106	2109	2111	2114	2117	2119	2122
1.63	2122	2125	2127	2130	2133	2135	2138	2140	2143	2146	2148
1.64	2148	2151	2154	2156	2159	2162	2164	2167	2170	2172	2175
1.65	2175	2177	2180	2183	2185	2188	2191	2193	2196	2198	2201
1.66	2201	2204	2206	2209	2212	2214	2217	2219	2222	2225	2227
1.67	2227	2230	2232	2235	2238	2240	2243	2245	2248	2251	2253
1.68	2253	2256	2258	2261	2263	2266	2269	2271	2274	2276	2279
1.69	2279	2281	2284	2287	2289	2292	2294	2297	2299	2302	2304
1.70	0.2304	2307	2310	2312	2315	2317	2320	2322	2325	2327	2330
1.71	2330	2333	2335	2338	2340	2343	2345	2348	2350	2353	2355
1.72	2355	2358	2360	2363	2365	2368	2370	2373	2375	2378	2380
1.73	2380	2383	2385	2388	2390	2393	2395	2398	2400	2403	2405
1.74	2405	2408	2410	2413	2415	2418	2420	2423	2425	2428	2430
1.75	2430	2433	2435	2438	2440	2443	2445	2448	2450	2453	2455
1.76	2455	2458	2460	2463	2465	2467	2470	2472	2475	2477	2480
1.77	2480	2482	2485	2487	2490	2492	2494	2497	2499	2502	2504
1.78	2504	2507	2509	2512	2514	2516	2519	2521	2524	2526	2529
1.79	2529	2531	2533	2536	2538	2541	2543	2545	2548	2550	2553
1.80	0.2553	2555	2558	2560	2562	2565	2567	2570	2572	2574	2577
1.81	2577	2579	2582	2584	2586	2589	2591	2594	2596	2598	2601
1.82	2601	2603	2605	2608	2610	2613	2615	2617	2620	2622	2625
1.83	2625	2627	2629	2632	2634	2636	2639	2641	2643	2646	2648
1.84	2648	2651	2653	2655	2658	2660	2662	2665	2667	2669	2672
1.85	2672	2674	2676	2679	2681	2683	2686	2688	2690	2693	2695
1.86	2695	2697	2700	2702	2704	2707	2709	2711	2714	2716	2718
1.87	2718	2721	2723	2725	2728	2730	2732	2735	2737	2739	2742
1.88	2742	2744	2746	2749	2751	2753	2755	2758	2760	2762	2765
1.89	2765	2767	2769	2772	2774	2776	2778	2781	2783	2785	2788
1.90	0.2788	2790	2792	2794	2797	2799	2801	2804	2806	2808	2810
1.91	2810	2813	2815	2817	2819	2822	2824	2826	2828	2831	2833
1.92	2833	2835	2838	2840	2842	2844	2847	2849	2851	2853	2856
1.93	2856	2858	2860	2862	2865	2867	2869	2871	2874	2876	2878
1.94	2878	2880	2882	2885	2887	2889	2891	2894	2896	2898	2900
1.95	2900	2903	2905	2907	2909	2911	2914	2916	2918	2920	2923
1.96	2923	2925	2927	2929	2931	2934	2936	2938	2940	2942	2945
1.97	2945	2947	2949	2951	2953	2956	2958	2960	2962	2964	2967
1.98	2967	2969	2971	2973	2975	2978	2980	2982	2984	2986	2989
1.99	2989	2991	2993	2995	2997	2999	3002	3004	3006	3008	3010

INDEX

A

Abscissa, 104
Accuracy, degree of, 402
  in a source, 34
Adjusted relatives, 343
Age of population in United States, 65
Aggregate, 22
  as an index, 278
Amplitude, 357
Analysis, of series, 91, 345
  statistical, 5
Annalist, The, 344
Arithmetic average of ratios, 279
Array, 193, 234
Arthur, H. B., 264
Automobile registration, 79
Average, 25
  arithmetic, 159
  exponential, 192
  for a frequency series, 168
  geometric, 161, 190
  good requisites of an, 157
  harmonic, 161, 190
  inverse-exponential, 192
  other simple, 280
  of price relatives, 273
  properties of, 155
  requisites of an, 157
  types of, 159
  weighted, 163
Average deviation, 200
  of a normal curve, 409
Axis, 103

B

Background for plotting curves, 107
Bank deposits, velocity of, 96
Bar diagram, 93
Base, price, 264
Bias, type, 290
  weight, 291
Bimodal, 135
Blank forms, questionnaire, 55
  for transcription, 77
Block diagram, 121, 128, 135
Blooming mills, employee earnings, 72
Bonds, issued by corporations, 376
  yield on, government, 372
Boots and shoes, earnings of employees, 124
  production of, 301, 346, 352, 387
    cycles in, 360
Business, number of concerns, 377
Business cycle, 361

C

Calculation of r, 241, 244, 245
Call loans, interest on, 383
Capacity of freight cars, 175
Capital of selected national banks, 140, 169
Captions, 70, 73
Categorical calculation of r from tables, 241
Categorical charting of series, 91
Categorical series, 13, 93
Cell, 234, 244
Cement production, 327, 331 341, 384
Central ordinate, 307
Characteristics, 127
  probable error of, 218
Chart, 15
  kinds of, 93
  publication, 92
  working, 92
Charting, categorical series, 91
  frequency series, 119
  instruction for, 405
  time series, 102
Cigars, tax-paid withdrawals, 382
Class interval, 119, 120, 174
Class limits, 119
Class mark, 120, 151

Class I railroads, freight cars, 89
operating revenues, 83
percentage, net to gross, 372
Classfication, rule of, 12, 62
Cluster, closeness of, 334
Code, coding of returns, 59
Coefficient, of correlation, 239
of dispersion, 206
quartile, 207
of skewness, 207
of variation, 207
Coke, production of, 88
Collection, defined, 38
Commodities, price of, 263, 286
Comparability of data, 36
Comparison of time graphs, 116
Component-part diagram, 95
Compound-interest curve, 318
Computation, instructions for, 400
of line of trend, 305
Concerns in business, number of, 377
Construction of tables, 60
Coordinates, 104
Corporate bonds, issued, 376
Corporate income, 371, 375, 376
Correlation, categorical data, 241
coefficient of, 239, 241
definition of, 230
derivation from table, 244
lagging, 363
multiple, 255
negative, 230
partial, 255
Pearsonian, 241
positive, 230
spurious, 254
table, description of, 231
time series, 363
Cotton, prices of, 266
production of, 31
yield of, per acre, 377
Cotton mill, wages of employees, 64, 68, 69, 380
Counting process, 64
Coupon rate on industrial bonds, 140
Cross moment, 241
Cross product, 241
Cross-section paper, 394
Crossing of index numbers, 290
Cumulative frequency, 135, 199
Curve, normal, 407
parameter of, 407
superimposed, 356
Curve-fitting, 186
by moments, 228
Cycles, 363
of boot and shoe industry, 360
of business, 361
of pig iron production, 356
in standard units, 358
of stock prices, 388
Cyclical fluctuations, 299, 353, 356, 363

D

Data, categorical charts for, collection of, 38, 50
primary, 50
revision of, secondary, 38, 50
Decile, 138
Deflation, statistical, 296
Deposits, national banks, 85
Determination of lag, 365
Deviation, average, 200
irregular, 299
from normal, 333
in normal curve, 409
quartile, 193
standard, 201
from trend, 350
Diagram, 15
bar, 16
block, 17, 121
component part, 95
pictorial, 97
scatter, 231, 234, 235
Discrete series, 15, 139, 175
Displacement between columns, 333
Dispersion, 127, 193
coefficient of, 206
measure of, 215
Dividends on stock, 180

E

Earnings, in blooming mills, 72
of bookkeepers, 176
in boot and shoe industry, 124
on common stock, selected, 232, 164

Earnings, of quarry employees, 180
Editing, 57
Elimination, of seasonal variation, 342
  of secular trend, 316
Enumeration, 53
Error curve, 208, 407
  of estimate, 131, 249
  probable, 215
  of sampling, 365
Estimated figures, 403
Excess reserves, of member banks, 376
Exponential average, 192
Exports, of gold, 372
  of iron and steel, 13
Extrapolation, 115

F

Federal Reserve Banks, deposits, of United States Treasury, 14
  excess reserves of, 376
  holdings of United States Government securities, 103
  loans, 11
  reserve ratio, computation of, 82
Fisher, Prof. Irving, 269
  his Ideal Index, 292
Fit, goodness of, 251
Fluctuation, cyclical, 299, 353
Footnotes, to indicate source and unit, 71
Forecasting, 363
  by analysis of cycles, 362
  sequence, 367, 368
Formally discrete, 175
Forms, blank, 55
  for computation, 80
  questionnaire, 55
  for transcription, 77
Frequency or frequencies, 119
  average of, 168
  charting, series, 119
  cumulative, 135, 199
  discrete, 139, 168
  grouped, 139
  moments of, 227
  polygon, 120, 135, 141
  weighted, 144
Frequency series, 12
  cumulative, 199
Frequency series, mean of, 174
  mode of, 185
  quartiles of, 198
Frickey, Prof. Edwin, 291, 368

G

Gaussian error curve, 209
General mean, 236, 240
Geometric average, 161, 190
  as index, weighted, 288, 289
Gold imports, 372
Goodness of fit, 251
Graph line, 102
Graphic examination of trend, 300
Graphic presentation, 92
Graphic representation of price changes, 267
Grid, 102, 104
Guides, in tables, 73

H

Harmonic index, weighted, 287, 289
Harmonic mean, 161, 190, 280
Harvard index, 388
Histogram, 17
Homoclitic, 237
Homogeneity, 18
  defects in, 34
  in frequency series, 35
Homoscedasticity, 237

I

Ideal Index, 292
Imports of gold, 372
Incomes, of corporations, 371, 373, 375, 376
  reported from dividends, 180
Increment rate, 307
Index numbers, 263
  adjusted for seasonal variation, 339
  aggregate, 278, 279, 281
  approach to problem, 277
  arithmetic (average), 281
  averages, 273, 279
  bias, 290, 291
  chain, 281
  of commodities, 269

Index numbers, crossing of, 291
economic conditions, 388
of factory employment, 160
fixed base, 281
general purpose, 285
geometric, 281, 288
harmonic, 281, 287
of industrial production, 42
of manufacturing, 160
median, 281
mode, 281
of prices, 147
purpose of, 269, 284
special purpose, 285
tests of good, 293
weights used by U. S Bureau of Labor Statistics, 373
Intercept, 137
Interpolation, 114
Interval, class, 13, 120, 139
Inverse-exponential average, 192
Iron and steel, 13, 14, 20
Iron-ore mines, 87
Isolated value, 139
Item, statistical, 7, 168

L

Labels, 7, 71
Laboratory procedure, 391
Lag, 362, 363
determination of, 365
maximum correlation, 363
Lead, 362
Least squares, principle of, 182
Legend, for maps, 95
Line graph, 102
Lines of regression, 247
Link relatives, 330, 335
median, 335
Logarithmic chart (ratio chart), 225
four-place table, 413

M

Maladjustment of prices, 277
Map, statistical, 94
Mean, 127, 155, 156
arithmetic, 155
Mean, general mean, 236, 240
geometric, 155, 280
harmonic, 155, 280
lines, 240
of price relatives, 277
properties of, 181
related to mode and median, 188
weighted, 280
Measures, of dispersion, 193, 215
of position, 199
Median, 138, 156, 159
of link relatives, 330, 335
properties of, 184
related to mean and mode, 188
Member banks, deposits of, 374
Mitchell, W. C., 273
Mode, 135, 156
of a frequency series, 185
as the maximum ordinate, 214
multimodal, 135
as related to the mean and median, 188
Moments, 221
cross (products), 241
first, second, third, 227
of a frequency series, 227
use in fitting curves, 228
Motor cars, registration by states, 79
Multimodal, 135
Multiple correlation, 255
Multiple frequency table, 133

N

National banks, deposits, 85
in Michigan, capital of, 140, 169, 379, 381
in Ohio, 378
National income, industrial origin, 373
New York Central Railroad, freight revenue, 30
Normal curve, areas under, 409
average deviation, 409
fitting of, 209
or law of error, 208, 407
ordinates of, 409, 412
quartiles of, 408
Normal points, 117
Normal trend, 351

O

Oats, price of, 266
Ogive, 135
Open ends, 244
Order of tabulation, 66
Ordering of items, 363
Ordinates, 104
  central, 307
  of normal curve, 214
  of trend, 311
Origin, 177
Original data, 38

P

Parameter, 209
Partial correlation, 255
Partial regression, 259
Pearsonian coefficient, 241
Per capita figures, 25
Percentiles, 138
Persons, Warren M , 291
Petroleum, production of, 383
Pictorial diagram, 97
Pie diagram, 100
Pig iron, production, 354, 385
    cycles of, 359
Polygon, of frequencies, 120
Population, by ages, of United States, 65
  statistical, 52, 130, 216
Portland cement, production of, 327, 341, 384
Price, base, 264
  changes of, 267
  commodity (index), 263
  copper, 109
  cotton, hogs, petroleum, pig iron, wheat, 286
  cotton, oats, wheat, 266
  graphic representation, 267
  maladjustment of, 277
  of selected commodities, 271
  of selected common stock, 232
Price index, 263, 269
  of the U. S. Bureau of Labor Statistics, 147
Price relatives, 264, 273
  frequency distribution of, 273, 274
Probability curve, 209
Probable error, 215
  of the characteristics, 218
Procedure, in laboratory, 391
Product, cross, 241
Production, of boots and shoes, 387
  of cement (Portland), 327, 331, 384
  of petroleum, 383
  of pig iron, 354, 385
Products, as variables, 25
Properties, of averages, 155
  of the median, 184
Publication tables, preparation of, 81
Puddling mills, wages of employees, 72
Punch cards, 64

Q

Quality, of a source, 46
Quantities, as weights, 289
Quarry employees, earnings of, 180
Quarterly data, analysis of, 345
Quartile, of a frequency series, 198
  of a normal curve, 408
  upper and lower, 138
Quartile coefficient, 207
Quartile deviation, 193
Questionnaire, 55, 56

R

Railroads, Class I, depreciation rate, 14
    freight cars, capacity, 89
    operating revenue, 83
    passenger miles, 89
Random sample, 216, 217
Range, 197
Rates, 23, 25, 166
Ratios, 23, 166
  chart, 319
  correlation, 253
  scale, 268, 328
Reference, to source, 10
Registration, 53
Regression, equations of, 258
  linear, lines of, 235, 247
  partial, 258
  planar, 260
Relative frequency, 123
Relatives, 263
  adjusted, 343, 351
  chain, 265 .

Relatives, fixed-base, 264
link, 264
median, link, 335
of prices, 264
Requisites of a good average, 157
Reserve ratio of Federal Reserve Banks, 82
Reserves, excess, for member banks, 376
Revisions, of data, 41, 403
Rulings, in tables, 72

S

Sales of two mail order houses, 108
Sample (sampling), 27, 51, 52
by design, 216
errors of, 131, 158, 215, 365
random, 216
representative, 216, 270
statistical, 130, 158
Scale, arithmetic, 16, 102, 108
logarithmic, 16, 108, 112
ratio, 108, 112
selection of, 104
Scatter diagrams, 231, 234, 235
Seasonal variation, 298, 326, 363
determination of seasonal indexes, arithmetic method, 340
logarithmic method, 338
deviation from trend, 350
distribution of link relatives, 332
elimination of, 342
division, 350
subtraction, 350
graphic evidence, 326
link relative method, 330
quarterly data, 345
tendency, 326
Secular trend, 298, 363
elimination of, 316
quarterly data, 347
Sequence, forecasting, 363, 367, 368
Series, 11–13, 18
categorical, 13, 18, 30, 93
discrete, 15
frequency, 13, 18
time, 13, 18
types of, 12
Silver production, 374
Skeleton method, 178
for computing, the coefficient of correlation, r, 245
Skeleton method, for computing, the mean, 177
the standard deviation, 205
Skew series, extremely skew, 223
moderately skew, 224
Skewness, 127, 221
coefficients of, 221
formulas, 222
Smoothing, 114, 129, 130, 132
of normal curve, 214
Source, 38
compilation from, 48
indicating, 71
original, 39
primary, 39
quality of, 45
reference, 10
secondary, 39
Spacings in tables, 73
Specification, of an item, 7
Spurious correlation, 254
Standard deviation, 201
in normal curve, 412
Standard units, 215, 358
in coefficient of correlation, 241
Statistical analysis, 5
Statistical arguments, 5
Statistical data, 5
Statistical items, 5, 7
Statistical map, 94, 97–99
Statistical series, 6, 11, 12
Statistics, 6
Stock prices, cycles of, 388
Stockholders of certain large corporations, 69
Stub, 12, 70, 73, 168
Summarization, analytical, 155
Summary numbers, 155
Superposed curves, 356
Survey, methods for conducting, 53
primary, 50, 51
Swings, cyclical, 356

T

Table of logarithms, 413
Tables, general, 60
presentation, 75
publication, 75, 81
summary, 61, 74, 75
working, 75, 76, 78
Tabulation, order of, 66

Tax returns, corporations, consolidated, 14
individuals, 28, 179
Tests, of a good average, 157
of a good index number, 291, 293
Time graphs, 116
Time series, 13
charting of, 102
components, 298
correlation of, 363
Ton-miles, percentage of net to gross, 372
Trade value of merchandise, 84
Transcription, specifications for, 391
Treasury bonds, 372
Trend, secular, 298
apparent, 304
broken, 317
compound interest, 319
curvilinear, 317
elimination of, 316
graphic examination of, 300
horizontal, 317
line of, computed, 305
negative, 317
ordinates of, 311
parabola, 322
selection of interval, 304
Type bias, 290
Types, of averages, 159
of series, 12
Typical value, 155

U

Unemployed, number of persons, 158
United States Bureau of Labor Statistics, wholesale price index, 373
United States Revenue collections, 373
United States Steel Corporation, number of employees, 29
United States Treasury bonds, 372
Units, 8
indicating, 71
lack of definiteness, 34
standard, 215, 358
as width of class interval, 177

V

Value of contracts awarded (index), 80
index of U. S. Bureau of Labor Statistics, 373
Values as weights, 285, 289
Variables, 18, 120, 168, 231
complexity of, 21
derived, 26
expressed in class intervals, 177
as products, 25
Variates, 18, 119
associated, 231
Variation, 3, 108
absolute, 163
coefficient of, 207
relative, 163
seasonal, 323, 298
elimination of, 342
Vehicles and equipment, production of, 86
Velocity of bank deposits, method of computing, 90

W

Wage rate, in building trades, 132
Wages, average for males under age 16, 81
of clerical workers, 122
of bookkeepers, 119
in cotton mills, certain employees, 380
Weighted averages, 163
Weighted frequencies, 144
Weighted indexes, aggregate, 289
arithmetic, 289
calculation of, 287
geometric, 288, 289
harmonic, 288, 289
Weighting, arbitrary, 284
explicit, 283
implicit, 283
index numbers, 283
purpose of, 284
Weights, calculation, 287
class, 295
commodity, 295
constant, 294
derivation of, 296
values in exchange, 285
Wheat prices, 266
Wholesale price index of U. S. Bureau of Labor Statistics, 373
Working table, purpose of, 78

Z

Zero points, 117