

Basic Statistics

A Textbook For The First Course

REVISED AND
ENLARGED EDITION

GEORGE SIMPSON & FRITZ KAFKA

A Textbook for the First Course

By GEORGE SIMPSON

Substantially enlarged and revised, this new edition of a highly successful textbook retains the approach of the original edition, clear, patient, and simple explanations of fundamental statistical concepts and procedures; nonmathematical treatment (only a knowledge of arithmetic and elementary algebra is needed); and practical emphasis on *application* of statistical methods.

The material on sampling has been expanded from one chapter to three and presents the elements of sampling theory, an explanation of statistical estimation and statistical significance, and a useful exposition of sampling practice. Valuable appendixes have been added on the normal curve, on rank correlation, and on trend ; the appendix on statistical quality control has been revised and expanded with an illustrative case history. The two chapters dealing with the function of statistics include additional material. The chapter on current indexes has been brought up to date. Many new graphs and tables have been added to the book.

U. S. Edition \$ 6.95

Third Indian Reprint ... Rs. 10,00



**INDIAN INSTITUTE OF
ADVANCED STUDY
LIBRARY SIMLA**

DATA ENTERED

CATALOGUED

Basic Statistics

BASIC STATISTICS

A TEXTBOOK FOR
THE FIRST COURSE

REVISED
EDITION

GEORGE SIMPSON

BROOKLYN COLLEGE

and FRITZ KAFKA

CHAS. PFIZER & CO., INC.



OXFORD & IBH PUBLISHING CO.

Calcutta

Bombay

New Delhi

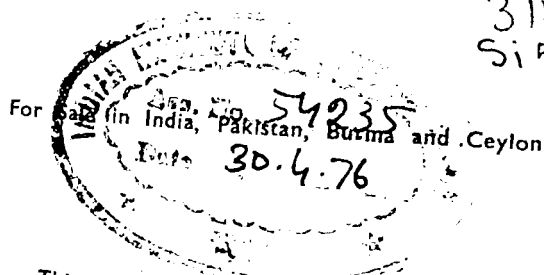
COPYRIGHT © 1952, 1957, BY
W. W. NORTON & COMPANY, INC.

INDIAN EDITION 1965

SECOND INDIAN REPRINT 1969
THIRD INDIAN REPRINT 1971

Rs. 10.00

311.2
5158B



This book has been published with the
assistance of the Joint Indian-American
Textbook Program

 Library IAS, Shimla



PL
17

00054235

PUBLISHING CO.,
BHUWANA-16 AND PRINTED BY
BHARAT LITHOGRAPHING CO., 98/4, S. N. BANERJEE
ROAD, CALCUTTA-14.

Contents

Prefaces	xvii
Acknowledgments	xx

PART 1. The Function of Statistics

1. THE IMPORTANCE OF STATISTICS	3
Statistics, Democracy, and Education	
The Scope of Statistics	
Statistics in Economics and Business	
The History of Statistics	
Qualifications of a Good Statistician	
Types of Statistical Work	
Statistical Thinking	
2. DATA—THE RAW MATERIAL OF STATISTICS	13
Data and Problems	
The Data in Statistics	
Collection of Data	
Presentation of Data	
Analysis of Data	
Interpretation of Data	
Inductive Statistics	
Truth and Statistical Data	

PART 2. Collection and Presentation of Data

3. COLLECTION AND SOURCES OF DATA	25
COLLECTION OF DATA	
Ways of Collecting	
Editing and Compiling the Data	
SOURCES OF DATA	
Types of Sources	
Leading Sources	

4. STATISTICAL PRESENTATION: TABLES	33
INFORMAL PRESENTATION	
Textual Presentation	
Semitabular Presentation	
TABLES	
Types of Tables	
Parts of Tables	
Construction of a Table	
How to Read a Table	
5. STATISTICAL PRESENTATION: LINE GRAPHS	41
ARITHMETIC LINE GRAPHS	
Elements of Arithmetic Line Graphs	
Constructing the Graph	
SEMILOGARITHMIC LINE GRAPHS	
The Logarithmic Scale	
Construction of a Semilogarithmic Graph	
Characteristics of a Semilog Graph	
Uses of Semilog Graphs	
Limitations of the Semilog Graph	
6. STATISTICAL PRESENTATION: GEOMETRIC FORMS, PICTURES, AND MAPS	65
GEOMETRIC FORMS	
Bar Charts	
Area Diagrams	
Volume Diagrams	
PICTOGRAPHS	
MAPS	
COMBINATIONS OF DIFFERENT TYPES OF GRAPHS	
COMPONENT-PART PRESENTATION	
Component Bar Chart	
Component Pictograph	
Component Line Graph	
Pie Diagram	
Choice of Component-Part Diagram	
FORMAL REQUIREMENTS	

MECHANICS OF GRAPHIC PRESENTATION

COMPARISON OF TABULAR AND GRAPHIC PRESENTATION

PART 3. Statistical Analysis

7. RATIOS 88

Meaning of Terms
Types of Ratios
Cautions Concerning Percentages
Some Important Ratios

8. THE FREQUENCY DISTRIBUTION 95

Raw Data

ARRAYS

The Simple Array
The Frequency Array

THE FREQUENCY DISTRIBUTION

Classes
Tally Sheet and Entry Form
The Frequency Table
Characteristics of Frequency Distributions
The Variable
Mid Point

PROBLEMS IN CONSTRUCTING A FREQUENCY DISTRIBUTION

Number of Classes
Actual Class Limits
Special Problems of Class Limits
Open-End Classes
The Actual Class Interval
Varying Class Intervals

PERCENTAGE FREQUENCIES

CUMULATIVE FREQUENCIES

9. TYPES OF FREQUENCY GRAPHS 115

Array Charts

GRAPHIC PRESENTATION OF FREQUENCY DISTRIBUTIONS

Histogram

Frequency Polygon
Different Shapes of Frequency Polygons
Ogive

10. MEASUREMENT OF MASSES: AVERAGES:
THE ARITHMETIC MEAN 127

The Concept of Average

THE ARITHMETIC MEAN

Mean from Ungrouped Data
The Weighted Mean
Mean from Grouped Data, Long Method
Mean from Grouped Data, Short Method
Characteristics of the Mean

11. MEASUREMENT OF MASSES: AVERAGES:
THE MEDIAN; THE MODE; THE GEOMET-
RIC MEAN 143

THE MEDIAN

The Concept of the Median
The Median for Ungrouped Data
The Median for Grouped Data
Characteristics and Uses of the Median
Related Positional Measures

THE MODE

The Concept of the Mode
The Mode for Ungrouped Data
The Mode for Grouped Data
Special Problems of the Mode

GRAPHIC ANALYSIS

Median and Related Measures
Mode

THE GEOMETRIC MEAN

Concept of the Geometric Mean
Computation of the Geometric Mean
Uses of the Geometric Mean
Limitations

12. MEASUREMENT OF MASSES: COMPARISON OF THE PRINCIPAL AVERAGES 163

1. Location of the Three Averages within the Frequency Distribution
2. Comparison of the Location of the Three Averages
3. Methods of Obtaining the Three Averages
4. Effect of Extreme Values on the Three Averages
5. Effect of Open-End Classes on the Averages
6. Varying Class Intervals and the Averages
7. Use of Averages in Further Computation
8. Mathematical Properties
9. Arrangement of Data and the Three Averages
10. Obtaining the Averages from Graphs

APPROPRIATENESS OF THE THREE AVERAGES

13. MEASUREMENT OF MASSES: VARIATION, SKEWNESS, KURTOSIS 173

VARIATION

Use of Measures of Variation or Dispersion

ABSOLUTE VARIATION: POSITIONAL MEASURES

The Crude Range

The Semi-Interquartile Range or the Quartile Deviation

The Quartiles for Ungrouped Data

The Quartiles for Grouped Data

Q and the Median

Characteristics of Q

ABSOLUTE VARIATION: COMPUTED MEASURES

THE AVERAGE DEVIATION

Average Deviation for Ungrouped Data

Average Deviation for Grouped Data

A.D. and the Average

THE STANDARD DEVIATION

Standard Deviation for Ungrouped Data: Long Method

Standard Deviation for Ungrouped Data: Short Method

Standard Deviation for Grouped Data: Long Method

Standard Deviation for Grouped Data: Short Method

THE NORMAL CURVE

The Standard Deviation and the Normal Curve

Use of the Standard Deviation

COMPARISON OF MEASURES OF ABSOLUTE VARIATION	
1. Type of Measure	
2. Relation to Averages	
3. Effect of Extreme Values	
4. Relation of Measures of Absolute Variation in a Normal Curve	
5. Relation of Measures of Absolute Variation to Algebraic Properties of the Averages	
6. Extent of Use	
RELATIVE VARIATION	
MEASUREMENT OF SKEWNESS	
Pearson's Measure of Skewness	
Bowley's Measure of Skewness	
KURTOSIS	
<i>APPENDIX—THE NORMAL CURVE</i>	210
Ordinates and the Normal Curve	
Areas under the Normal Curve	
14. INTRODUCTION TO TIME-SERIES ANALYSIS	221
What is Time-Series Analysis?	
ELEMENTS OF TIME SERIES	
Trend	
Seasonal Variation	
Cyclical Variation	
Irregular Variation	
PREPARATION FOR ANALYSIS OF A TIME SERIES	
Editing Time-Series Data	
Graphic Presentation of Data	
15. TREND	235
Reasons for Trend Analysis	
THE MEASUREMENT OF TREND	
DETERMINING TREND BY INSPECTION OR ESTIMATE	
The Freehand Method	
The Selected-Points Method	
DETERMINING TREND BY COMPUTATION: THE SEMI-AVERAGE METHOD	

DETERMINING TREND BY COMPUTATION: THE LEAST-SQUARES METHOD	
Introductory Illustration	
Least-Squares Long Method	
Least-Squares Short Method	
USE OF THE TREND EQUATION	
Shifting the Origin	
THE METHOD OF MOVING AVERAGES	
Limitations of the Moving-Average Method	
ADJUSTMENT FOR TREND	
CURVILINEAR TREND	
<i>APPENDIX—SPECIAL PROBLEMS OF TREND ANALYSIS</i>	264
CONVERSION OF ANNUAL TREND EQUATION TO MONTHLY TREND EQUATION	
Where Data Are Annual Totals	
Where Data Are Given as Monthly Averages per Year	
Time Values in Half-Yearly Units	
Shifting the Origin	
NONLINEAR TREND BY LEAST SQUARES	
16. SEASONAL VARIATIONS	272
Reasons for Measuring Seasonal Variations	
The Specific Seasonal and the Typical Seasonal	
Computation of Seasonal Variations	
ADJUSTMENT FOR SEASONALITY	
17. CYCLICAL AND IRREGULAR VARIATIONS; FORECASTING	288
THE PROBLEM OF CYCLES	
Statistical Characteristics of Cycles	
MEASURING CYCLES BY THE RESIDUAL METHOD	
Annual Data	
Monthly Data	
IRREGULAR FACTORS	
FORECASTING	
Importance of Forecasting	

Methods of Forecasting	
Procedures in Statistical Forecasting	
Limitations of Statistical Forecasting	
18. INDEX NUMBERS	302
Importance of Index Numbers	
Index Numbers and Other Statistical Concepts	
Classification of Index Numbers	
PROBLEMS IN THE CONSTRUCTION OF PRICE INDEX NUMBERS	
Data	
Base	
Combining the Data	
Weighting	
SPECIAL PROBLEMS	
Making Indexes Comparable	
Combining Index Numbers	
Splicing	
Percentage Change in Index Numbers	
QUANTITY INDEX NUMBERS	
VALUE INDEXES	
SPECIAL-PURPOSE INDEXES	
19. CURRENT INDEXES	329
IMPORTANT PRICE INDEXES	
Wholesale Price Index	
Consumer Price Index	
Other Price Indexes	
IMPORTANT QUANTITY INDEXES	
The Federal Reserve Board's Index of Industrial Production	
World Index of Industrial Production	
VALUE INDEXES	
SPECIAL-PURPOSE INDEXES	
20. INTRODUCTION TO CORRELATION	348
The Concept of Correlation	
Correlation and Causation	

Spurious Correlation	
The Scatter Diagram	
Types of Relationship	
BASIC CONCEPTS	
The Regression Line	
Standard Error of Estimate	
Coefficient of Correlation	
Coefficient of Determination	
COMPUTATION OF MEASURES	
Computation of r	
Computation of S_y	
Computation of the Regression Line	
SCOPE OF CORRELATION	
<i>APPENDIX—RANK CORRELATION</i>	375

PART 4. Inductive Statistics

21. ELEMENTS OF SAMPLING THEORY	381
Sampling in Everyday Living	
Induction	
THE UNIVERSE AND THE SAMPLE	
Why a Sample Is Used	
SAMPLE-UNIVERSE RELATIONSHIPS	
Concepts of Estimation and Statistical Significance	
THE THREE DISTRIBUTIONS	
The Universe Distribution	
The Sample Distribution	
The Sampling Distribution	
Relations of the Three Distributions	
INTERPRETATION AND USES OF THE STANDARD ERROR OF THE MEAN	
STANDARD ERROR OF OTHER STATISTICS	
Standard Error of the Median and of the Standard Deviation	
Standard Error of the Total	
Standard Error of a Proportion	

Standard Error of a Difference
Standard Error and Sample Size
Standard Error and Universe Size

22. ESTIMATION AND SIGNIFICANCE 413

ESTIMATION

The Concept of Estimation
Estimation of the Mean
Estimation of Other Measures
Estimation and Sample Size

STATISTICAL SIGNIFICANCE

The Concept of Statistical Significance
The Null Hypothesis
Difference between Sample Mean and Universe Mean
Difference between Sample Proportion and Universe Proportion
Difference between Means of Two Samples
Difference between Two Sample Proportions
Limitations of Tests of Significance

23. SAMPLING PRACTICE 430

RANDOM SAMPLING

Random Selection
Restricted Random Samples
Stratification
Cluster Sampling
Systematic Sampling
Other Sample Designs
Sample Size

PURPOSIVE SAMPLING

Comparison of Purposive and Probability Sampling

OTHER TYPES OF SAMPLING

The Chunk
Sequential Sampling

APPENDIX—STATISTICAL QUALITY CONTROL 450

CONCEPT AND HISTORY

Acceptance Sampling
Process Control

CONTROL CHARTS

- The Concept
- Types of Control Chart
- How Control Charts Are Constructed
- How to Read a Control Chart
- An Illustration of the Use of a Control Chart

APPENDIX—RANDOM NUMBERS AND THEIR USE 461

PART 5. Misuse

24. MISUSES OF STATISTICS 465

The Problem of Misuse

MISUSES IN THE COLLECTION OF DATA

- Incomparable Data
- Failure to Consider Changes in Classification
- Biased Sample
- Incomplete Enumeration

MISUSES IN THE PRESENTATION OF DATA

- Failure to Present Complete Classification System
- Spuriously Accurate Presentation
- Errors in Graphic Presentation

MISUSES IN THE ANALYSIS OF DATA

- Use of Absolute Numbers Instead of Percentage
- Use of Percentages Instead of Absolute Numbers
- Faulty Use of Percentages
- Misuse of the Mean
- Failure to Use the Weighted Mean
- Faulty Use of the Median
- Faulty Use of the Mode
- Faulty Use of the Range
- Failure to Use a Measure of Dispersion
- Faulty Extrapolation of Trend
- Faulty Use of Indexes
- Misuse of Correlation

MISUSES IN THE INTERPRETATION OF DATA

- Failure to Comprehend the Total Background of the Data

Interpretation Based on Individual Cases Instead of Average	
Interpretation Based on Average Instead of Individual Cases	
Confusion of Averages	
Interpreting Seasonal Variation as Cyclical Variation	
Interpreting Cyclical Variation as Seasonal Variation	
Interpreting Seasonal Variation as Trend	
Time Sequence Interpreted as Causation	
Misinterpretation of the Coefficient of Correlation	
CONCLUSIONS ON MISUSES	

General Appendixes

I. HOW TO MAKE A STATISTICAL REPORT	483
II. APPROXIMATE NUMBERS AND ROUNDING	486
III. HOW TO TAKE A SQUARE ROOT	490
IV. TABLE OF SQUARES AND SQUARE ROOTS	493
V. TABLE OF COMMON LOGARITHMS	505
VI. BIBLIOGRAPHY	508
INDEX	514

Preface to the Revised Edition

In revising *Basic Statistics* we have remained steadfast in our aim to emphasize thoroughness at the introductory level and not to include material that cannot be handled at that level. Even so, we have been able to introduce some additional material at the suggestion of the teachers who received the first edition so cordially.

Changes in this revised edition are the following:

1. We have expanded the material on sampling from one chapter in the first edition to three chapters in this revision. This expansion has permitted the statement of the elements of sampling theory, explanation of statistical estimation and statistical significance, and a useful exposition of sampling practice (including discussion of the types, size, and design of samples).
2. We have completely rewritten the appendix on statistical quality control and have placed it at the end of the last chapter on sampling (Chapter 23). (In fact, we have placed all substantive appendixes at the ends of their respective chapters in this revised edition, rather than at the end of the book with more general material and numerical tables.)
3. We have added an appendix on the normal curve, indispensable for understanding sampling theory, to Chapter 13, Measurement of Masses: Variation, Skewness, Kurtosis.
4. We have added an appendix on rank correlation to Chapter 20, Introduction to Correlation.
5. We have reworked Chapter 3, Collection and Sources of Data, and we have omitted material on questionnaire construction which appeared in the earlier edition as too specialized and out of place in an introductory statistics text.

6. We have brought up to date Chapter 19, Current Indexes, **taking** into consideration the revisions of the Wholesale Price Index, of the Consumer Price Index, of the Index of Industrial Production, and of the Index of Department Store Sales.

7. We have expanded the appendix on trend to include the fitting of second-degree parabolas by least squares, and we have placed it at the end of Chapter 15, Trend.

8. We have substituted many new graphs and tables for those in the first edition, bringing data up to date whenever doing so would serve a useful purpose.

9. We have added material to Chapters 1 and 2 to strengthen these chapters on orientation.

10. We have prepared a separate book of Problems and Exercises to replace those that were at the back of the text in the first edition; this Problems and Exercises book includes needed work materials.

11. We have brought the bibliography up to date.

We have been aided, in the new sampling chapters especially, by the close readings and learned advice of our former colleagues: Mr. John E. Murphy of Bristol-Myers Products Division, and Mr. Irving Roshwalb of Audits and Surveys Company, Inc. Mr. Paul Chook, also a former colleague, now of Alfred Politz Research, Inc., gave us the benefit of his extensive knowledge of statistical quality control in his reading of the appendix on that subject. Dr. Benjamin Lipstein of Audits and Surveys Company, Inc., read the sampling chapters and made valuable and important suggestions.

Mrs. Josephine Harkay brought her skill to making the new drawings for charts and illustrations in this revised edition.

As before, we depended on Willis L. Parker of W. W. Norton & Company for expert editorial help and he did not fail us.

George Simpson
Fritz Kafka

Preface to the First Edition

This textbook is especially designed to help the beginning student master the grammar of statistics, but does not purport to be a handbook for the statistical practitioner. The content of the book has been carefully fitted to this aim, and no more has been included than we believe will achieve our goal.

Basic Statistics is intended for the elementary, one-semester course in statistics on the undergraduate level, or for use in the first semester of a longer course, to lay the foundation for more advanced work. Thus this book should prove suitable as a text either for terminal elementary training or as a basis, and we hope a stimulant, for further work in statistics.

To meet the need for laboratory work within one cover, we have supplemented the text with problems and exercises tied in with the content of the chapters. Such material gives the student the opportunity to round out his initial training by applying the concepts to practical problems.

Statistics is here dealt with as a living, human activity, not an abstract body of formulas and procedures. Hence, we have tried in every chapter to connect technical aspects with statistical thinking and practice. Moreover, the discussion has been restricted to the basic concepts. Thoroughness, not all-inclusiveness, has been the aim.

Since undergraduate students are usually required to understand only elementary algebra and arithmetic in order to begin the study of statistics, this book presupposes no more than these. We have therefore stressed the logic of formulas and methods in nonmathematical terms. Frequent use has been made of visual illustrations to explain abstract concepts. Furthermore, on the basis of our joint experience in teaching the introductory course in statistics, we have tried to anticipate the questions usually asked by students.

We have sought to make a contribution to statistical pedagogy on the elementary level and therefore have assumed the

risks always connected with simplification of sometimes involved procedures; but we think the advantages more than compensate for the risks. These advantages should include ease of communication and readability. In keeping with this effort, we have sometimes used hypothetical data to illustrate techniques and conjectural situations to help explain statistical measures and concepts.* The laboratory problems and exercises, however, involve actual data from government, industry, and business.

In the interest of stressing underlying concepts, where several methods have been developed for the same purpose we usually have selected just one. The instructor can present alternative methods if he wishes, with, we hope, some assurance that the student is now familiar with the concepts involved.

Thus, here and throughout, the aim has been to enable the instructor to build upon the text by freeing him as far as possible from the need to explain basic elements.

In the chapters on sampling and correlation, we have sought to give a rounded introduction to these wide fields that can still be handled in the one-semester beginning course. The chapter on current indexes describes the recent developments of government bureaus regarding the major indexes.

In the appendixes we have included, in addition to the usual aids, a simple statement on the preparation of a statistical report. For elementary practical work in sampling, we have also included in the appendixes a table of random numbers with a description of their use, and a short discussion and illustration of control charts used in statistical quality control.

Acknowledgments

It is difficult to state what we owe to other authors of books in this field. We must have assimilated a great deal from them that

* Whenever we have used hypothetical data, they have been treated as if collected by ourselves and therefore no source note is given.

we can no longer separate from our own understanding. We are indebted to all of our professional colleagues whose contributions to statistics and the teaching of statistics have helped us.

Harold S. Diamond of Queens College gave us a very careful reading of the manuscript, and his suggestions were of great value. Sol Dutka, statistician of Dun and Bradstreet, Inc., made important recommendations concerning the chapter on sampling, and also gave us his ideas on the book as a whole. Irving Roshwalb, statistician of the Opinion Research Corporation of Princeton, New Jersey, went over the entire manuscript and made very important criticisms. Paul H. Chook, Fellow in Statistics at the City College of New York, read the manuscript with particular regard to its usefulness to beginning students, and gave good advice. Benjamin Lipstein of the United States Bureau of Labor Statistics read the galley proofs and gave us his expert opinion.

For an intelligent job in typing the manuscript we are indebted to Miss Jean Leventhal. For technical aid we wish to express thanks to our student, Jerome H. Skolnick. In the preparation of the manuscript for the printer, Willis L. Parker carefully went over the entire book with us and gave us valuable suggestions and advice which improved our final formulations. Miss Helen E. Nielsen made the illustration drawings. Finally, for encouragement and counsel we wish to express our gratitude to the late Addison C. Burnham of W. W. Norton & Company and to Arthur D. Gayer, Chairman of the Economics Department of Queens College.

We alone, however, are responsible for any errors that may appear. Both of us have contributed equally to this book and share responsibility for its contents.

PART

1

The Function of Statistics

CHAPTER

1

The Importance of Statistics

Statistics, Democracy, and Education

A citizen faces a barrage of statistics in his newspaper, in his magazines, in advertising, over television and radio, and in books. He must seek to penetrate these numerical mysteries; his citizenship in a democracy demands a participation that can be intelligent only if he can appraise and evaluate quantitative information. In a democracy, "the citizen lives in a world of facts and figures. He makes decisions all of the time on the basis of large or small amounts of information. He carries on in a mass-production economy. . . . Perhaps H. G. Wells was right when he said 'statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.'" *

To deal with statistics one needs to be trained in statistics. A grasp of statistics has thus become an educational "must," a part of what is called "general education" which aims to "bring the student into an awareness of and harmony with the statistical content of our society." †

* S. S. Wilks, "Undergraduate Statistical Education," *Journal of the American Statistical Association*, March 1951, vol. 46, No. 253, pp. 1-18.

† George W. Snedecor, "A Proposed Basic Course in Statistics," in *Journal of the American Statistical Association*, March 1948, pp. 53-54.

Though the American Statistical Association was founded in 1839, popular knowledge of statistics remained limited for many years. In 1887, Carroll D. Wright, United States Commissioner of the Bureau of Labor, wrote of the necessity for the State and Federal governments to "be vitally interested in the elevation of statistical work to scientific proportions; for the necessary outcome of the application of civil service principles to the conduct of all government affairs lies in this, that as the affairs of the people become more and more the subjects of legislative regulation or control, the necessity for the most accurate information relating to such affairs and for the scientific use of such information increases." *

Today, "to a very striking degree our culture has become a statistical culture. Even a person who may never have heard of an index number is affected in an intimate fashion by the gyrations of those index numbers which describe the cost of living. Even on the most elementary level it is impossible to understand psychology, sociology, economics, finance, or the physical sciences without some general idea of the meaning of an average, of variation, of concomitance, of sampling, of how to interpret charts and tables. The deliberations of Congress and the state legislatures deal continually with matters in which it is impossible to reach a sound decision without weighing statistical evidence." †

The Scope of Statistics

The word *statistics* today refers either to quantitative information or to a method of dealing with quantitative information. In the first reference it is used as a plural noun—"the statistics of production in this company are as follows"; in the second reference the word is used in the singular—"statistics deals with the collection, presentation, analysis, and interpretation of quantitative information."

* Carroll D. Wright, "Statistics in Colleges," *Publications of the American Economic Association*, vol. III, No. 1, March 1888, p. 25.

† Helen M. Walker, "Statistical Literacy in the Social Sciences," *The American Statistician*, February 1951, pp. 6-12.

Statistics pervades all subject matters. In their meetings, professional statistical societies discuss such topics as "Statistics in Housing Research and Planning," "Industrial Accident Statistics," "Censuses of Population and Agriculture," "Quantitative Measures of Efficiency in Marketing," "Statistical Methods in Highway Traffic," "Business Statistics: the Stock Market Picture," "Employment Statistics," "Educational Testing," "The Statistics of Industrial Management," "Statistical Quality Control," "Statistical Methods in Astronomy," "The Statistics of Marriage and Divorce," "Statistics in Biology, Chemistry, and Physics." Statistics is thus a tool of all science, indispensable to research and intelligent judgment. It has become a recognized discipline in its own right.

Statistics in Economics and Business

The fundamental concepts of statistics are the same in all fields but these concepts are emphasized and utilized somewhat differently in each field. In economics and business, certain statistical concepts gain importance because of the subject matter. Though this book is concerned chiefly with statistics in economics and business, it is well to remember that in economics and business we occasionally meet problems of statistical application that are associated with other fields—for example, problems of psychological and educational statistics arise in personnel administration.

It would be difficult to overestimate the importance of statistics to an understanding of business, industry, and labor problems, to the workings of government, and to the study of economic processes. The twentieth century has seen a growth in statistical application that would utterly astound a citizen of the nineteenth century. Thus recently it has been said that no economist would attempt to arrive at a conclusion concerning the production or distribution of wealth without an exhaustive study of statistical data.* The intervention of government in the economy,

* Carl C. Engéberg in "The Statistical Method," *Encyclopedia Americana*, vol. 25, p. 529, 1951.

the growth of large-scale entrepreneurial activity, the introduction of scientific methods into various parts of business administration, the growth of mass organizations of workers and consumers—all have stimulated and contributed to the rapid development of economic and business statistics in the twentieth century.

The History of Statistics

The recent flourishing does not mean that statistics has no early history. Censuses of population and wealth were taken by the Pharaohs and the ancient Hebrews. According to the Greek historian Herodotus, Rameses II in 1400 B.C. took a census of all the lands of Egypt in order to reapportion territory. We have similar reports on the ancient Chinese, on the Greeks, and on the Romans. People and land are thus the earliest objects of statistical inquiry.

Although the word *statistics* was used before the eighteenth century, it appears that Gottfried Achenwall in 1749 was the first to use the term to refer to a subject matter as a whole. Achenwall defined statistics as "the political science of the several countries." The so-called "university statistics" in Germany and what became known in England as "political arithmetic" are the two great tributaries of the stream which became modern statistics.

Great mathematicians of the eighteenth and nineteenth centuries helped pave the way for modern statistics. Here belong the names of Bernoulli on probability theory and least squares, of Gauss on the normal curve and least squares, and of Quételet on the discovery and interpretation of variability.

The term *statistics*, up until the last quarter of the nineteenth century, was used to signify not only numbers and quantitative information but also facts calculated to illustrate the conditions and prospects of society. But by the turn of the twentieth century the term *statistics* became identified with quantitative

information and today this is almost the exclusive emphasis which is given to the subject.*

We approach the contemporary scene with Sir Francis Galton and Karl Pearson. Pearson's name is inextricably connected with the development of modern statistical theory, and several statistical devices bear his name. Further advances, indispensable to contemporary statistical theory, grew out of the original and outstanding work of R. A. Fisher.

So great has been the influence of statistical method that, as a recent president of the American Statistical Association has written, "although statistics is in its infancy in a scientific sense, it is certain to influence profoundly all future scientific thinking." †

Qualifications of a Good Statistician

Clearly the technical details of statistical measurement must be grasped in order to understand quantitative information and interpret it correctly. But these technical details are not enough. What other equipment must a trained statistician have? An answer must be given in terms of (1) knowledge ‡ and experience, and (2) personality.

The statistician who applies statistical methods to a subject matter must have familiarity with the subject matter in addition to technical skill in the handling of figures. For example, a statistician in industry needs to know the intimate and intricate details of the industry, its production methods, its history, its customary practices, its economic problems, its reporting system and sources of information, and the like.

"Good judgment, broad knowledge and experience, and

* Walter F. Willcox in "Statistics: History," *Encyclopedia of the Social Sciences*, vol. xiv, p. 357.

† Lowell J. Reed, "Man as a Planning Animal," *Journal of the American Statistical Association*, vol. 47, no. 257, March 1952, p. 4.

‡ See *Educational Requirements for Employment of Statisticians*, VA Pamphlet 7-8.9, United States Government Printing Office, 1955. Prepared by the Bureau of Labor Statistics, U. S. Department of Labor.

common-sense are the most valued possessions of the statistician and research worker." *

The ideal personality traits that make for a good statistician may be summarized in the words of the Institute for Research of Chicago, Illinois: "Those who work with statistics . . . must be accurate and painstaking. Indeed they must have a passion for accuracy. There is no place in the field for a slovenly worker." †

In addition to having these qualities, a good statistician is a person of imagination and improvisation. Practical situations require deft adaptation of statistical techniques to the problems at hand. Slavish adherence to the letter will not suffice; the spirit, here as elsewhere, lifteth up.

Types of Statistical Work

The qualifications cited are necessary to the statistician who *applies* statistics. Not all of them are indispensable to one who devotes himself solely to the *mathematical* side of statistics. Indeed, we may distinguish four types of statistical worker: (1) the mathematical statistician; (2) the applied statistician; (3) the statistical administrator; (4) the statistical assistant. The United States Civil Service Commission classifies statisticians falling in the first three groups here from a slightly different viewpoint. They distinguish mathematical, analytical, and survey statisticians.

The mathematical statistician is interested in working out the abstract theory of statistical method. He concerns himself also with developing new techniques. Consequently, this type of statistician must have a thorough knowledge of advanced mathematics and its application to statistics. Though sometimes thought of as removed from practical statistical concerns, mathematical statisticians have latterly realized the need to keep in touch with the problems of the statistical practitioner.

* Robert E. Chaddock, *Principles and Methods of Statistics*, Houghton Mifflin Company, p. 31.

† "Statistical Work as A Career," Chicago, 1946.

On the other hand, the applied statistician must understand the findings of the mathematical statistician although he is not expected to arrive at them himself. As his designation indicates, the applied statistician puts statistical methods into practice in a particular subject matter; thus we speak of business statisticians, economic statisticians, social statisticians, biostatisticians, educational statisticians. As a recent United States government publication has put it: "Since the intelligent application of statistical techniques to the study of specific problems requires a sound knowledge of the field in which the study is being made, most statisticians must be well-trained in the particular subject-matter fields in which they use their statistical skills: for example, biology, public health, agriculture, economics, sociology, psychology, engineering, and market or other business problems. Applied statisticians usually remain in their own special field or in related fields of study, because, in their case, knowledge of the subject-matter is usually as important as knowledge of statistical methods." *

The statistical administrator is the supervisor of the collection and presentation of statistical data. He is in charge of machine tabulation, editing, charting, and similar functions. Although not a statistical analyst, he must nevertheless understand the work of the applied statistician.

The category of statistical assistant includes clerks, typists, draftsmen, enumerators, and computers. These compose a very large group. Statistical training is usually not required of them, but some knowledge of statistical problems is desirable.

From the point of view of statistical practice, we may helpfully distinguish between general practitioners in statistics and specialists. The growth and development of statistics has brought about this division of labor, and one function of a general practitioner that is emerging involves coordinating the work of specialists much as does the general practitioner in medicine.

* *Employment Outlook in the Social Sciences*, "Statisticians," Bulletin No. 1167, United States Department of Labor, Bureau of Labor Statistics, in cooperation with the Veterans Administration, 1954, p. 17.

Statistical workers of the various types are employed in many different fields. The *Roster of Scientific Personnel*, a publication of the United States government, lists the major sources of employment in statistics in order of their importance as follows:

Government (Federal, State, and local); manufacturing firms; banks, insurance companies, and financial institutions; public utilities and railroads; social agencies; business, labor, and other types of national associations; educational and research institutions; wholesale and retail trade organizations; advertising and market-research firms. In government they work in labor, welfare, health, highway, agricultural, taxation, banking and insurance, and education agencies.

Statistical Thinking

Statistical thinking is concerned with the quantitative characteristics of a mass of items and differences within this mass; for example, the total labor force in the United States comprises a mass of items and these may be differentiated according to occupation, age, sex, income. There is variation in each of these aspects, and the study of such variations is a main concern of statistical thinking. None of these aspects is invariant throughout the mass. Nor does statistics study any of these aspects in terms of any single individual item. The unqualified application to the individual item of the findings obtained from masses is sin against statistical thinking. The subject matter of a statistical study is thus not a particular object but the entire collection of objects distinguished by certain properties.* What statistics is concerned with is the variation in characteristics of masses. Consequently, making comparisons between masses or within a mass is a basic activity of the statistician; indeed, statistics has been called the art of comparison. In our illustration of the labor force, for instance, we could compare incomes at one time period or age distribution at different time periods.

Statistical thinking differs from historical thinking in so far as the latter is concerned with unique objects, persons, or events,

* Oskar N. Anderson, "Statistical Method," in *Encyclopedia of the Social Sciences*, vol. xiv, p. 367.

such as *the* Panama Canal, *the* Governor of California, *the* Battle of the Bulge. In contrast to those parts of natural science that seek universal or invariant relations, statistical thinking is in terms of probabilities, approximations, and averages. But statistical thinking is not foreign to the natural scientist; he talks in terms of probabilities when he cannot establish invariant relations. But certain parts of natural science deal with uniformities in objects, persons, or events; an example is the law of gravitation. Where the uniformity of the laws of nature has been challenged, the statistician looks for "laws obeyed only on the average by large aggregates of individuals; so he takes as his province the study of the behavior of such aggregates. . . . The statistician . . . resigns himself to the impossibility of predictions with astronomical accuracy; but he tries to measure how often his predictions will go astray." *

Statistical thinking is a form of logical thinking, and is a part of our daily lives. When we say something or somebody is typical, we are thinking in terms of statistical averages, and departure from type is statistical variation. When we generalize from a few cases to a very large number, we employ sampling processes. A conclusion that two things always go together involves a pattern of thinking found in statistical correlation. Hence, statistical thinking is not foreign to everyday thinking but is a scientific form of it. Statistics is a fundamental activity of mankind.

Summary

1. Democratic citizenship requires knowledge of statistics by the citizen. It is a part of general education.
2. Statistics is a tool for all scientific research.
3. In economics and business the twentieth century has witnessed a tremendous growth in the use of statistics.

* Maurice G. Kendall, "The Content of Statistics," speech delivered as part of the Bicentennial Celebration of Columbia University, New York City, May 1954.

4. Statistics has a long history, but its rapid development dates back only to the recent past.

5. Four types of statistical work have emerged as distinct occupational groups in the field. These are the work of the mathematical statistician, the applied statistician, the statistical administrator, and the statistical assistant.

6. Statistical thinking may be distinguished from thinking in history and natural science, in so far as history deals with the individual occurrence and the sciences may be concerned with the uniformities and invariant relations, whereas statistics is concerned with variations and differences.

CHAPTER

2

Data—the Raw Material of Statistics

Data and Problems

All thinking—and this obviously includes statistical thinking—begins with a problem. A problem involves a felt difficulty; it is necessary for us to act—but how? Problems are rooted in the necessity for making decisions, and clear decisions can only be made in terms of evidence.

The aimless accumulation of quantitative data may be of service to anyone who wishes to surprise or impress by sudden demonstration of unusual knowledge, but it is not any part of statistics. Statistical work must be directed toward actual or potential problems. Data have no standing in themselves; they have a basis for existence only where there is a problem.

Thus, statistics does not properly concern itself with amassing numerical information in the hope that it may be useful to solve problems. Statistics is concerned with amassing data in order to solve problems, and even where there is collected a vast assemblage of figures arrived at by what seems a “figure factory,” this assemblage is presumably designed to aid in the solution of specific problems.

The Data in Statistics

Not all quantitative data are statistical. Isolated measurements are not statistical. Data are statistical when they relate to measurement of masses, not statistical when they relate to an individual item or event as a separate entity. The wage earned by an individual worker at any one time, taken by itself, is not a statistical datum; taken as part of a mass of information, it may be a statistical datum. Thus all the wages earned in the plant or industry in which the individual works, or in the occupation in which the individual is involved, or in the geographical area in which he resides or works, may be statistical data. Moreover, the wages earned by one worker over a period of time, being a series of wages, can be used statistically.

Though statistics deals with quantitative data, for purposes of interpretation and action the quantitative data may not be enough; they may need to be complemented by historical data, descriptive data, or knowledge gained through other non-quantitative sources. The wider the vision and learning of the statistician, the more apt is he to see significant relationships in the data he examines.

Collection of Data

In statistical work, the first step is to secure data.

But in statistics, as in all scientific pursuits, the investigator often need not *begin from the very beginning*; he may use, and must *take into account*, what has already been discovered by others. Consequently, before starting a statistical investigation we must read the existing literature and learn what is already known of the general area in which our specific problem falls, and any and all surrounding information that may give us leads and lessen the number of pitfalls and unnecessary labors and duplications of effort.

When research is done on a problem where the data have not yet been assembled (for instance, in such fields as public-opinion research, market research, and other types of research where the

data must be amassed, as it were, on the spot), it is necessary for us to go out and collect information ourselves in terms of the *specific* problem at hand.

When collecting data we must know what we are talking about; the terms we use must be unambiguously defined. To take an elementary example, if we are going to investigate wages we must decide whether we mean annual wages, monthly wages, weekly wages, daily wages, or hourly wage rates. Fuzzy or casual definition of terms will continually harass the investigator, and may result in the collection of data that are not comparable at all.

We shall deal in Chapter 3 with general problems of collection of data, and in Chapter 23 with specific problems of collection in the case of sampling.

For data to be reliable they must be collected by sound methods. Statistical results can never be better than the data upon which they are based. Moreover, even if the data have been collected by rigorous standards and techniques, failure to handle them correctly, as by making mathematical errors, corrupts the end product.

Quantitative data, particularly when they are presented in complicated fashion, are often so impressive that people accept them as their own justification for being. But unreliable data can be manipulated by the same statistical methods as reliable data.

Everyone realizes that we should measure the phenomena which we treat whenever we can, and that increasing precision in measurement is a scientific gain. But there is danger that the seductions of statistical technique may blind enthusiasts to the imperfections and inadequacies of the data.*

If there has been no standardized, uniform method of recording each and every individual item which makes up the mass we are studying, the results are worse than useless; they are misleading. Or if there have been no uniform instructions rigorously carried out by the enumerators or those making the measurements, the

* W. C. Mitchell, "The present status and future prospects of quantitative economics," *American Economic Review*, XVIII, pp. 39-41, March 1928.

figures are worthless. Unreliability in the original data renders all further statistical manipulation and interpretation utterly meaningless.

The responsibility for the reliability of statistical data is generally placed on the collector, but reliability also should concern the user. A leading economist has indeed complained that "as a rule, collectors and publishers of primary data do not deem it their obligation to accompany a series by a detailed description of how it was obtained; and users also, for the most part, tend to accept a series, particularly one issued by a governmental agency, at its face value without inquiring into its reliability." *

Presentation of Data

After data have been collected, they can be presented. Statistical data may be presented informally or in tabular or graphic form. A table can give a very accurate presentation; it can offer the actual figures. A graph, which presents quantitative data in visual form, ordinarily gives only a more or less close approximation. From careful and painstaking reading of tables and graphs, pertinent and revealing facts are discoverable. What may take pages of text to say can be said briefly in tabular and graphic presentation.

Presentation of data will be discussed in Chapters 4, 5, and 6.

Analysis of Data

Sometimes presentation is an end in itself, and sometimes it is intertwined with analysis.

Once reliable data have been collected on a mass basis, we can then classify them, condense them, summarize them, correlate them, isolate the elements of a composite force, depending upon the data and what we are studying. In some instances, analysis can be done graphically.

1. CLASSIFICATION OF DATA. We may have data on incomes for instance, or on textiles, or on production of copper. But within

* Simon Kuznets, "Conditions of Statistical Research," *Journal of the American Statistical Association*, vol. 45, No. 249, March 1950, p. 12.

each of these broad categories we can make subdivisions that will advance our knowledge and increase our insight into the data. For example, we might want to classify incomes according to their source, as wages, dividends, profits, or rents; textiles according to their kind—cotton, wool, rayon and other synthetics, silk; copper production by years and countries.

This systematic breakdown of the data may be sufficient for certain analytic purposes or may be preparation for further manipulation.

In making a classification, we must keep it consistent throughout. And wherever possible, such classifications as "etc." and "miscellaneous" should be avoided.

Sometimes a system of classification involves more than one aspect. We may need a breakdown of the data according to different attributes. For example, we may want to know the classes of wages in terms of numerical limits (as \$50.00 to \$60.00 a week), and also to know how these wages are distributed in terms of different occupations. These occupations might be classified broadly (for instance: white-collar, industrial, agricultural, self-employed, managerial), or they might be broken down still further into particular functions (for instance: clerks and typists, machinists and helpers, farm laborers and food processors, retail merchants and professionals, executives and administrators).

But classification may be only a step toward further analysis.

2. **CONDENSATION AND SUMMARIZATION.** The most important method of the condensation and summarization of data involves the use of what we call a frequency distribution and the finding of averages and related measures. These measures describe, and are representative of, the entire mass of items with which we originally started.

3. **CORRELATION.** In correlating data we seek to show how a mass of items is related quantitatively in its ups and downs to the ups and downs of another mass of items with which it is connected.

4. **ISOLATION OF ELEMENTS.** In analyzing data over time, we are seeking to isolate recurrences and trends. This type of

analysis breaks the data down into what we call cycles, seasonality, irregularities, and long-term tendencies.

5. GRAPHIC ANALYSIS. We mentioned before that graphs are a means of presenting data. In addition, graphs can sometimes be used to establish certain averages, to indicate correlation, and to analyze data over time. Such graphic analysis may be sufficient to solve our problem, or may be a valuable preliminary step to more refined analysis.

Analysis of data will be discussed extensively in Part III.

Interpretation of Data

Of the four parts of statistical work—namely, collection, presentation, analysis, and interpretation of data—the last is the area of least agreement. But there are four fundamental principles of interpretation concerning which everybody will agree.

1. *Sound interpretation involves willingness on the part of the interpreter to see what is in the data.* That is, there are no interests greater than truth. Statistics are too often interpreted to prove what is euphemistically called “policy”; having decided upon what is to be proved, the interpreter works the figures over to prove it. There is no *statistical* answer to such unwillingness to heed the facts, particularly when backed by power to put this heedlessness into practice.

2. *Sound interpretation of statistics requires that the interpreter know something more than the mere figures.* He must be fully aware of the problem and background to which the statistics pertain. He must have a thorough and systematic knowledge of the whole subject matter, an understanding of the relation of the subject matter to allied bodies of knowledge, and an intimate and special familiarity with the problem at hand.

3. *The rules of logical thinking are indispensable to sound interpretation of statistical data.* Logical thinking keeps the statistician from fallacious interpretation. The abilities to arrive at correct conclusions from premises, to reason inductively and deductively, have no substitutes.

4. *Clear, incisive language is part of sound interpreting.* The

choice of language is determined by the level of the prospective users. The student should seek to learn how to communicate his interpretations to one not trained in statistics.

Part V of this book deals with an aspect of interpretation.

Inductive Statistics

The summary above has dealt with methods used to *describe* data: the area of statistics often called *descriptive statistics*. Other methods are needed when we wish to *generalize* from the data we have to the larger group that the data represent. This latter area, known as *inductive statistics*, is dealt with in Part IV of this book and is of utmost importance in present-day statistics.

Truth and Statistical Data

Despite the high prestige that statistics enjoys and its indispensable use in scientific investigation, there remains among the general public an undercurrent of sentiment to the effect that "you can prove anything by statistics," or that statistics is only window dressing for conclusions reached on quite other grounds, or that the same figures used by different people lead to different conclusions. This view is sometimes expressed by the statement that statistics gives the appearance of exactitude to statements concerning things we really do not know much or anything about.

Some may want to use figures to sway opinion in the direction of their vested interest, some may want to attract attention by seeing in data a relationship that is not there. All sorts of pressures are applied to workers in statistics, just as pressures are applied to politicians, reporters, critics, and the like. Moreover, statistics has been known to be subjected to human prejudice. Prejudice involves unwillingness to abide by the weight of evidence. It means deciding what you are going to discover regardless of the statistical data.

The statistician, however, need not be swayed by pressure and prejudice. To be sure, the statistician, very often not being his own master, may see his findings used for ulterior purposes, and

may even be asked to "angle" them. But the fact that he works in a world where other interests may clash with the truth should not deter the statistician from his high calling: it should merely give him greater insight into the way of a world in which there may be conflict between conformity and the standards of scientific performance.

The high calling of the statistician was the theme of Carroll D. Wright, pioneer in the establishment of statistics in higher education and in government in the United States, who late in the nineteenth century wrote words which still ring true concerning the application of statistics to social and economic problems:

If there is an evil, let the statistician search it out; by searching it out and carefully analyzing statistics, he may be able to solve the problem. If there is a condition that is wrong, let the statistician bring his figures to bear upon it; only be sure that the statistician employed cares more for the truth than he does for sustaining any preconceived idea of what the solution should be. A statistician should not be an advocate, for he cannot work scientifically if he is working to an end. He must be ready to accept the results of his study, whether they suit his doctrine or not. The colleges in this connection have an important duty to perform, for they can aid in ridding the public of the statistical mechanic, the man who builds tables to order to prove a desired result. These men have lowered the standard of statistical science by the empirical use of its forces.*

Summary

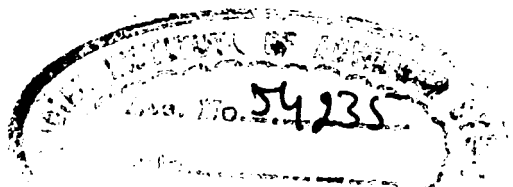
1. The presence of a problem gives meaning to statistical data.
2. Statistical data give quantitative information about masses, and frequently must be supplemented by nonquantitative information for full understanding of the data.

* Carroll D. Wright, "Statistics in Colleges," *Publications of the American Economic Association*, vol. III, No. 1, March 1888, p. 27.

3. Unreliable data make useless and even dangerous all statistical work that proceeds from them.

4. The collection, presentation, analysis, and interpretation of data make up four steps in statistical work. A distinction is usually made between inductive and descriptive statistics.

5. Statistical data may be used incorrectly because of human frailties, but this possibility is no reflection upon statistics as a scientific tool.



PART

2

Collection and Presentation of Data

CHAPTER

3

Collection and Sources of Data

How and where shall we get statistical data? We may collect them ourselves or take them from available sources. Sometimes we have no choice and must collect information ourselves, either because none is available in the form we need or because data relating to our problem are not sufficiently reliable.

Time and expense are often crucial factors in our decision whether to collect data or to take them over. In general, collection is a comparatively expensive and time-consuming procedure; frequently a large staff must be employed for this purpose.

COLLECTION OF DATA

Ways of Collecting

If we are to collect the data ourselves, how do we go about it? When we collect data ourselves, we do so through what we may call *direct investigation* as distinguished from investigation through sources. In direct investigation we may obtain data either through observation or through inquiry. In *observation* the collection is, as it were, one-sided; for example, when we

measure the lengths of certain machine parts or count the number of people passing a given show window at different hours of the day.

In *inquiry* we ask people questions. These questions may be asked through a personal interview or by a mail questionnaire. On occasion answers to an inquiry may be obtained through having people register information. Sometimes we combine methods of inquiry.

In *personal interviewing* questions are asked either face-to-face or by telephone. Personal contact is absent in *mail questionnaires*. If we have a choice in collecting data either through personal interview or through the mail, on what grounds do we make a decision as to which one to use? Below are some of the advantages of each method of data collection.

Personal Interview. (1) Aimed at specific respondents; no "self selection" by respondents. (2) Large response rate; that is, high percentage of returns. (3) Permits explanation of questions concerning difficult subject matter. (4) Permits evaluation of respondent, his circumstances, and his reliability. (5) Useful where spontaneity of response is required. (6) Personal rapport may help to overcome reluctance to respond. (7) Permits probing to explore questions in depth. (8) Promptness of returns—no "dribbling in."

Mail Questionnaire. (1) No possible influencing of respondent by interviewer. (2) Mailing costs much lower than costs of personal visit. (3) Geographically dispersed respondents can be quickly reached. (4) Respondents can be reached without appointment or concern for when they will be available. (5) Permits respondent to remain anonymous. (6) Reaches all groups, including those where personal solicitation is not possible. (7) Time available where considered response is necessary.

The leading problems in constructing interview schedules and mail questionnaires have been classified * under the following four headings:

* Arthur Kornhauser, "Constructing Questionnaires and Interview Schedules" in *Research Methods in Social Relations*, The Dryden Press, New York, 1951, Part II, pp. 423-462.

1. Decisions regarding Question Content.
2. Decisions regarding Question Wording.
3. Decisions regarding Form of Response to the Question.
4. Decisions about the Place of the Question in the Sequence.

In a basic text on statistics we cannot discuss exhaustively the construction and use of interview schedules and mail questionnaires. These two techniques of collection are nevertheless of great importance in practical work. A specialized literature has grown up since the middle of the 1940's, which can be found in the bibliography in this book under the heading "Survey Techniques" on page 510.

Editing and Compiling the Data

After the data have been collected through observation or inquiry, we must prepare them for presentation and analysis. The mass of raw material comes in, as a rule, without any systematic arrangement: a pile of questionnaire or interview forms appears on the statistician's desk.

It has become established practice to have a trained editor check over the forms or other returns for completeness and consistency. The editor may be able to facilitate later work by making corrections wherever necessary. By appropriate inquiries, he may be able to salvage a form that otherwise would have to be discarded. It is not always necessary to edit every form. A sample from the mass may be sufficient to appraise the returns.

Frequently, it will greatly simplify this work if answers are translated into a simple code. If sales territory, for instance, has been recorded as one element of the investigation, we may break down the total territory into parts. For example, New England might be designated as 01, the Middle Atlantic States as 02, the South Atlantic States as 03, and so on. Or if size of locality is of importance, we may code M for metropolitan area (cities over 100,000 population), LC (large city) for cities of 50,000 to 100,000, and so on.

For mechanical tabulation, coding of each answer from each form is required since such tabulation machines as those of

IBM 770371 IBM 142396
COMMODITY CARD

REFERENCE		ENTRY	DATE	BRANCH	SALESMAN	CUSTOMER	INDUSTRY	QUANTITY	UNIT	COST PRICE	SALES PRICE	COST AMOUNT	SALES AMOUNT	UNIT SALES PRICE
COMMODITY							12	DZ						
COMMODITY CODE CUSTOMER NUMBER INDUSTRY CODE STATE CITY BRANCH SALESMAN REFERENCE NUMBER ENTRY DATE UNIT														COMMODITY UNIT PRICE COST SALES
COMMODITY UNIT PRICE COST SALES														

Illustration 3.1. Punch Card for Sales Analysis, International Business Machines Corporation.

International Business Machines (IBM) or Remington Rand are constructed to record only coded answers. The codes may be holes punched in cards. A hole must be punched in a special place on a card for each answer. A key-punch machine punches the hole at the correct spot in an appropriate column. (See Illustration 3.1.)

Where the use of machines is not called for, the information is transposed onto what is called a tally sheet. (See Illustration 3.2.)

Type of Movie	Age and Sex			
	Under 10		10-15	
	M	F	M	F
Musical	②	### I ⑥	### ### ⑩	### ### ### ⑬
Western	### ### ### ⑩	③	### ### ### ⑫	②
Mystery			### ### I ⑪	
Comedy	### ⑤	### ### ### ⑮	### ### ⑩	### ### ### ⑮
Melodrama			②	### ### ### ⑭
Documentary			I ①	①

Illustration 3.2. Tally Sheet Showing Motion-Picture Preferences by Age Groups and Sex in Mill City for 160 School Children under 15.

After the compact and systematic assemblage of data on punch cards or tally sheets is completed, the next task is to summarize the information. Data on tally sheets may be totaled direct; if data have been transferred to punch cards these are mechanically sorted and then totaled. Mechanical sorting is a quick and efficient process wherein the cards are passed through the machine, which separates them into set groups according to their characteristics as punched on the cards.

When the processes of editing and compilation have been completed, the data are ready for presentation and analysis.

This short description of editing and compilation is not to be taken to mean that these tasks are necessarily short and unimportant. No step in a statistical inquiry can be taken lightly. If editing and compilation are not done competently, presentation and analysis will be of no value.

SOURCES OF DATA

The statistician's data may be collected directly and specially for him, as described and discussed in the section preceding. But he may choose to use data already collected and developed by others; such statistical information may be entirely applicable to the problem he is considering. The persons or organizations that have gathered the data, and the reports or publications in which the data are published, are then the sources of the data. For instance, a wealth of statistical information is contained in publications of government agencies, trade and industry associations, research organizations, and in certain periodicals and newspapers. The use of statistics to guide governmental action and economic enterprise has become widespread, and statistical reporting is being increasingly emphasized.

Types of Sources

Sources of data are referred to as primary or secondary. A primary source is one that itself collects the data; a secondary source is one that makes available data which were collected by some other agency. The files of a trade association or its publications are illustrations of primary sources. If we take trade-association data from the *Wall Street Journal*, then the *Wall Street Journal* is a secondary source for these data. A primary source usually has more detailed information, particularly on the procedures followed in collecting and compiling the data. A secondary source is not, however, necessarily an inferior source; in much practical work, a secondary source is just as acceptable as a primary source. It must be noted that a given

source may be partly primary and partly secondary. The Labor Department's *Monthly Labor Review*, for instance, uses data compiled by the Labor Department as well as data compiled by the Commerce Department and other federal agencies.

Leading Sources

In the United States, the federal government is the largest supplier of economic and business statistics. Such agencies as the Bureau of the Census and the Bureau of Labor Statistics, the Bureau of Agricultural Economics, the Bureau of Mines, the National Office of Vital Statistics, the Securities and Exchange Commission, the Interstate Commerce Commission, and many others offer us a continuing, regular flow of statistical information. So widespread and varied are the statistical activities of federal departments and bureaus that an Office of Statistical Standards (in the Bureau of the Budget) has as its chief function the study of the coordination of all this statistical work. State and local governments, in varying degrees, also provide such information. The United Nations has become a leading source of statistical data.

But there are also very important nongovernmental sources. Trade and industry associations collect data from members and publish much of this material. Large corporations and labor unions have statistical departments. Private research organizations are also important. Here belong such organizations as the National Bureau of Economic Research, the National Industrial Conference Board, and Dun and Bradstreet. Trade papers, economic journals, and some newspapers are also sources of data.

The student should early become acquainted with such publications as the *Statistical Abstract of the United States*;^{*}

^{*} The *Statistical Abstract of the United States*, published annually since 1878, is the standard summary of statistics on the industrial, social, political, and economic organization of the United States. It is compiled, edited, and published by the Bureau of the Census. It includes a representative selection of data from most of the important statistical publications, both governmental and private. Emphasis is given primarily to national data. The *Statistical Abstract of the United States* has grown from 157 pages in the 1878 edition to more than 1000 pages in 1956.

the *Survey of Current Business*, which is published monthly by the Bureau of Foreign and Domestic Commerce of the Department of Commerce; the *Monthly Labor Review*, published by the Bureau of Labor Statistics of the United States Department of Labor; the *Federal Reserve Bulletin*, published monthly by the Federal Reserve Board; and other such publications. A useful guide to government data has been published by the Office of Statistical Standards of the United States Bureau of the Budget; its title is *Statistical Services of the United States Government*. A nongovernmental publication of like value is *Government Statistics for Business Use*, by Hauser and Leonard.

Having collected data on our own or obtained data from a source, we have laid the foundation for statistical inquiry and are ready for presentation, analysis, and ultimately interpretation. The better the foundation, the sounder the structure that may be built on it.

Summary

1. Statistical data are obtained, in direct investigation, through observation or inquiry.
2. The data collected through direct investigation must be edited and compiled.
3. Compilation can be done by machine (key punch and sorting) or by hand (tally sheet).
4. Data for a statistical investigation may be taken from sources that have collected them. The sources may be primary or secondary.
5. The leading source of economic and business statistics is the United States government. There are also important nongovernmental sources.

CHAPTER

4

Statistical Presentation: Tables

Good presentation of statistical data is not always an end in itself; it frequently sets the stage for analysis of the data. Through good presentation, significant facts and comparisons are highlighted, and attention to them leads to intelligent use of the statistical information.

Statistical information may be presented without formal organization, in a formal table, or in graphs. In this chapter we shall deal with informal and tabular presentation. Graphs will be dealt with in Chapters 5 and 6.

INFORMAL PRESENTATION

Textual and semitabular presentations are considered informal.

There is no need for any set of rules for the elementary textual or semitabular forms of presentation.

Textual Presentation

In a discussion of steel production, for example, statistics can be made part of the running text; thus:

The American Institute of Steel Construction reported that estimated total bookings of fabricated structural steel for January 1950

amounted to 116,987 tons. This compares with 124,251 tons booked in December and 130,418 tons booked in January 1949.

This type of presentation is not to be used for a large amount of information.

Semitabular Presentation

This consists of setting off the figures in the text discussion, usually by indenting and sometimes by change of type. For example, the figures quoted directly above may be presented semitabularly thus:

The American Institute of Steel Construction reported estimated total bookings of fabricated structural steel as follows:

January 1950.....	116,987 tons
December 1949.....	124,251 tons
January 1949.....	130,418 tons

This semitabular arrangement is also called "leader work." Its advantage over textual presentation is that it brings the figures closer together and thus makes comparisons easier.

TABLES

A table is a systematic organization of statistical data in columns and rows. Rows are horizontal arrangements; columns are vertical. The purpose of a table is to simplify the presentation and to facilitate comparisons. In general, the simplification results from the clear-cut and systematic arrangement, which enables the reader to quickly locate desired information. Comparison is facilitated by bringing related items of information close together.

Types of Tables

The basic types of tables are the reference (or general-purpose) table, and the text (or special-purpose) table. The reference table is a repository of information whose purpose is to present

detailed statistical material. Many complete United States Census tables are reference tables. On the other hand, text tables have an analytical purpose. They bring out a specific point or answer a specific question.

Accordingly, reference tables are usually far larger than text tables. They are found in appendixes of publications, or as parts of general compendiums of information. Text tables, however, accompany the pertinent textual discussion.

It follows from the different characteristics of these two types that the arrangement of the reference table should aim at ease of reference, whereas the text table should emphasize items, relationships, or comparisons of significance to the specific problem.

Parts of Tables

Certain parts must be present in all tables. There are other parts whose presence depends upon the specific case.

The parts that must be present are as follows: (1) title, (2) stub, (3) caption (or box head), (4) body (or field). Other parts, that may be present, are: (5) headnote (or prefatory note), (6) footnote, (7) source note.

1. A complete title, which appears at the top of the table, has to answer the questions *what*, *where*, and *when* in that sequence.* These are necessary in order to fully describe and delimit the contents, and to guide the reader to the information he desires. A good title is compact, yet complete. If the complete title proves unwieldy, it may be preceded by a short "catch" title.

2. The stub consists of the stub head and the stub entries. The stub head describes the stub entries, whereas each stub entry labels the data found in its row of the table.

3. The caption (or box head) labels the data found in the columns of the table. The caption consists of one or more column heads. Under a column head there may be subheads.

* Sometimes the title also states *how* the data are classified.

4. The body (or field) contains the numerical information.
5. A headnote (or prefatory note) is a phrase or statement below the title which clarifies the contents of the table or main parts of it; for instance: *All data in long tons.*
6. A footnote is a phrase or statement which clarifies some specific item or some specific part of the table, or explains the

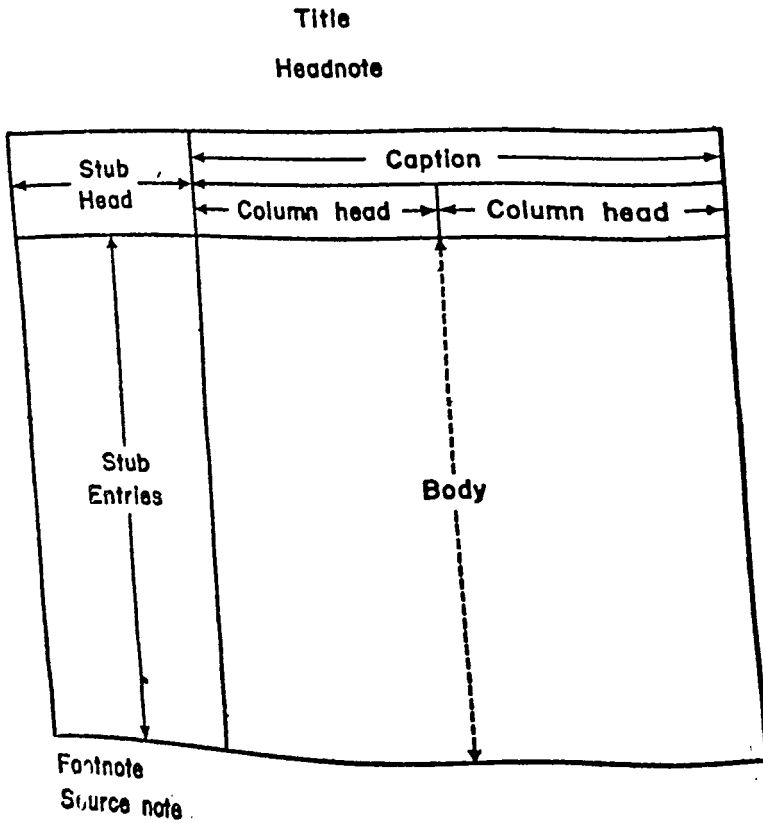


Illustration 4.1. Format of a Table.

omission thereof, and is placed at the bottom of the table; for instance: *The figure for 1957 is an estimate.*

1. A source note is used to state clearly where the data were obtained if they were not collected by the one presenting them. It is exceedingly important to state the source, for this permits

the reader to check the figures and possibly gather additional information. Moreover, it is part of professional ethics to give credit where credit is due. For these reasons, the source note has to be unambiguous, and complete as to title, edition, time, page, and sometimes place of publication.

A schematic diagram of a table with parts labeled is shown in Illustration 4.1.

Construction of a Table

There are no hard and fast rules on constructing tables. But the ability to construct a good table is not as easily acquired as may at first be thought.

To show system in tabular presentation, we must arrange the data in keeping with the purpose of the presentation and the nature of the data. This arrangement may be alphabetical, geographical, or by magnitude, to mention a few possibilities.

Eleven guides for table construction are as follows:

1. Certain places in the table give stress. Thus, if we wish to emphasize the total of a column of figures, we place the total at the top of the column.
2. Do not plan the size and shape without a preliminary layout that shows that all the data to be presented can be accommodated and arranged properly. Keep in mind that the table may have to appear in print, and adjust the size and shape accordingly.
3. The title of a table is placed at the top, and is centered.
4. If the rows must be very long, then the stub should be repeated at the right.
5. Indicate a zero quantity by a zero, and do not use zero to indicate that information is not available. If it is not available, show this fact by the letters *n.a.* or by a dash (—).
6. Abbreviations should be avoided, especially in titles and headings. For example, "Yr." should not be used for "Year."
7. Be explicit. The expression "etc." or "et cetera" is bad form in a table, since the reader may not readily discover what it refers to.

8. Do not use ditto marks. If a figure is repeated, show it each time. A ditto mark may be mistaken for the figure 11.

9. In headings, whenever possible, use the singular; a column of products should be headed "Product" and not "Products."

10. Be consistent in ruling. For instance, if ruling is used to close a table on one side, use it also on the other side.

11. If footnotes are keyed to numbers in the table, avoid using numbers to designate the footnote. Use symbols, such as asterisks and daggers, whenever possible.

Some of these guides are not universally applicable or universally accepted. In some cases it will be impossible to follow them. But keep their purpose in mind. This discussion may be aptly summarized in the words of a former director of the United States Bureau of the Census, J. C. Capt: "In the final analysis, there are only two rules in tabular presentation that should be applied rigidly: *First*, the use of common sense when planning a table, and *second*, the viewing of the proposed table from the standpoint of the user. The details of mechanical arrangement must be governed by a single objective; that is, to make the statistical table as easy to read and to understand as the nature of the material will permit." *

How to Read a Table

We have been discussing the production of tables, and now a few words are in order on the consumption of tables. Let us examine a table that appeared in the *Monthly Labor Review* for February 1947 (facing page).

This example is a text table. What we are looking for in a text table are comparisons and relationships among the data. These comparisons and relationships can be established by following items through columns or rows.

Looking at the title, we see that the table is a study of the agricultural labor force in 1945 and the changes that have

* U. S. Bureau of the Census, *Bureau of the Census Manual of Tabular Presentation*, by Bruce L. Jenkinson, U. S. Government Printing Office, Washington, D. C., 1949, p. iii.

TABLE 1. FARM WAGE WORKERS IN 1945 CLASSIFIED BY EMPLOYMENT STATUS IN JANUARY 1946¹

Employment status in January 1946 of persons who worked for wages on farms in 1945	Both sexes		Males						Females	
			Total		Less than 150 days worked ²		150 or more days worked			
	Number (thousands)	Percent	Number (thousands)	Percent	Number (thousands)	Percent	Number (thousands)	Percent	Number (thousands)	Percent
Total farm wage workers..	3,212	100	2,375	100	1,657	100	718	100	837	100
In the labor force.....	2,279	71	2,035	86	1,329	80	706	98	244	29
Employed in agriculture.....	1,640	51	1,548	65	924	56	624	87	92	11
Employed in nonagriculture.....	563	18	413	18	346	20	67	9	150	18
Unemployed.....	76	2	74	3	59	4	(³)		(³)	
Not in the labor force.....	933	29	340	14	328	20	(³)		593	71

¹ Source: U. S. Department of Agriculture, Bureau of Agricultural Economics, Employment and Wages of the Hired Farm Working Force in 1945, June 1946; mimeographed. Estimates relate to persons 14 years of age and over in the civilian noninstitutional population in January 1946, who did farm work for wages in 1945.

² Includes 9,000 males for whom information on number of days worked was not obtained.

³ Estimated as less than 50,000.

occurred in this labor force in January 1946. The title also indicates that these changes can be seen according to employment status. A footnote leads us from the title to the source, which we thus find to be a mimeographed publication of the United States Department of Agriculture, Bureau of Agricultural Economics. Ordinarily, such information is found in a source note not linked to the title as a footnote. Moreover, the use of numbers to designate footnotes is not recommended.

The stub groups the 1945 labor force in terms of employment status in January 1946, breaking down the total into those in the labor force in January 1946 and those not in the labor force in January 1946. Those still in the labor force in January 1946 are broken down still further into those employed in agriculture, those employed in nonagriculture, and those unemployed.

Looking beyond the stub to the first pair of columns, we can see that of the total farm wage workers numbering 3,212,000 in 1945, only 2,279,000 or 71% were still in the labor force in

January 1946. The remaining portion, 933,000 or 29%, had left the labor force.

This table also permits us to compare male and female agricultural workers. Of the original 3,212,000 workers in 1945, 2,375,000 were men and 837,000 were women. Of the men, 340,000 or 14% had left the force in January 1946, as compared with 593,000 women or 71% of the women. There is an additional breakdown of the number of male workers who worked more or less than half a year in 1945. In general, following the columns permits comparison among the employment groups of the farm wage workers, while following the rows enables us to compare for each employment group the absolute and percentage figures, the share of part-time and full-time male workers, and the share of males and females in that employment group.

This discussion should suffice to show the procedure to be followed in reading a table.

Summary

1. Statistics may be presented without formal organization, in a table, or in a graph.
2. Tables may be classified as reference tables or text tables.
3. Certain parts must be present in all tables (title, stub, caption, body); other parts may be present (headnote, footnote, source note).
4. There are some basic guides for planning and constructing a table.
5. In reading a text table, we look for comparisons and relationships by following items through columns and rows.

CHAPTER

5

Statistical Presentation: Line Graphs

A graph (or chart) is a visual form for presentation of statistical data. The main types of visual presentation are (1) lines (or curves), (2) geometric forms, (3) pictures, or (4) maps. Line graphs may be arithmetic, semilogarithmic, or logarithmic. The geometric forms include rectangles (or bars), circles (or pies), squares, and cubes, and thus may be one-dimensional, two-dimensional, or three-dimensional. Statistical charts consisting of pictures are called pictographs. Maps are used to present statistical data geographically.

In this chapter we shall discuss arithmetic and semilogarithmic line diagrams. Geometric forms, pictures, and maps will be discussed in Chapter 6.

There are special types of graphs that are reserved for the presentation of some one field of statistics. These include frequency graphs and scatter diagrams. We shall discuss the first of these in connection with frequency distributions in Chapter 9, and the second in connection with correlation in Chapter 20.

ARITHMETIC LINE GRAPHS

Elements of Arithmetic Line Graphs

A line graph presents, as a line or curve, variations of related magnitude against the background of what is called a grid. There are two elements of variation, one called the independent variable, the other called the dependent variable. In time series, time is always the independent variable; the quantity of crops produced in a series of years, or sales made in a series of months, might be the dependent variable. The independent variable is measured along a horizontal axis called the X -axis. The dependent variable is measured along a vertical axis called the Y -axis. The Y variable is called dependent because it *depends upon* the independent (or X) variable; that is, once a value for X is fixed, the corresponding Y value is determined. In our illustration, time is placed on the horizontal axis and sales on the vertical axis. The horizontal axis is also called the abscissa, and the vertical axis the ordinate.

The horizontal axis is drawn, and at a point marked zero on this axis a perpendicular is erected which is designated as the Y -axis. The zero point is also known as the point of origin. The field called Quadrant I in Illustration 5.1 is used when both variables are positive values, which in practical work is almost always the case. If we extend the two axes, we obtain three more quadrants which are used if either variable or both contain negative values. If losses of a corporation over time are shown as negative profits, for instance, we would have to use Quadrant IV.

Along the horizontal and vertical axes are placed scales. These scales start at the point of origin. Successive unit distances are laid off from the origin along each axis, and thus the scales are established. At equal distances on the X -axis we draw lines parallel to the Y -axis, and at equal distances on the Y -axis we draw lines parallel to the X -axis. These lines are called grid lines and the background they form for a graph is called the grid. They may form squares or other rectangles.

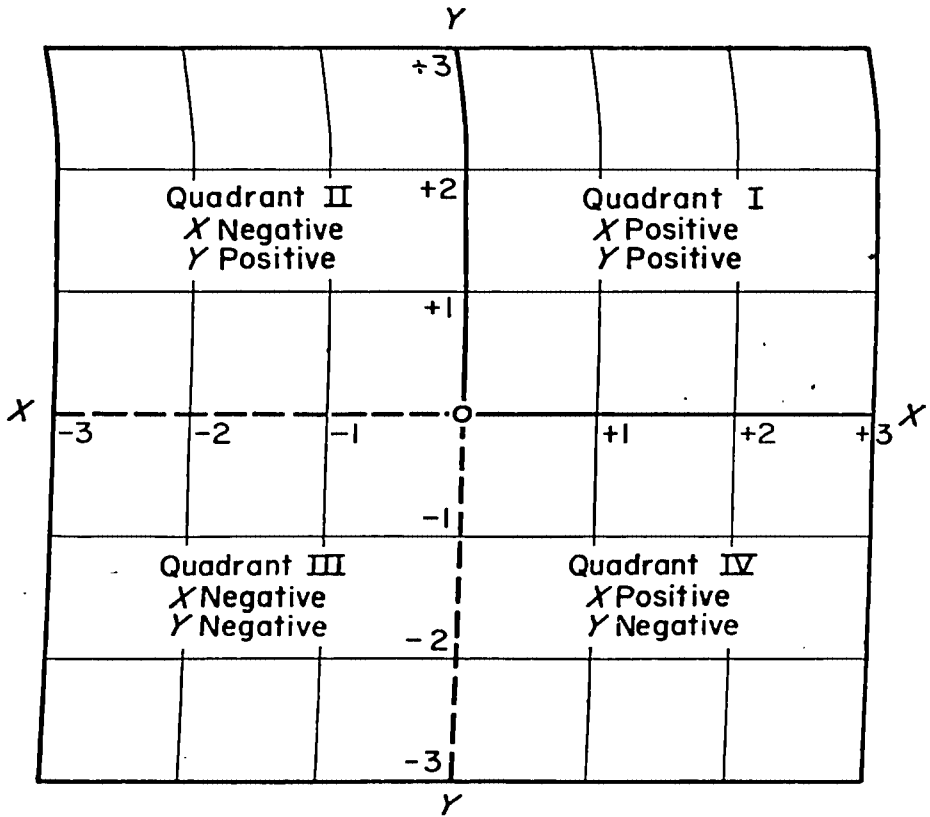


Illustration 5.1. Coordinate System for Plotting Arithmetic Graphs Showing Axes, Quadrants, Grid Lines, and Scales.

Each scale on this kind of graph is in *arithmetic* progression, hence the terms *arithmetic* line graph and *arithmetic* grid.

Constructing the Graph

There are seven rules or guides to follow in constructing an arithmetic line graph.

1. *Zero Origin.* On an arithmetic line graph, the *Y* scale must begin at zero as origin. Thus the *X*-axis always runs through this zero origin, the zero line is the base line, and the curve is interpreted in terms of distance from this base line. In one special case, when we present graphically a series of

changes from a norm of 100%, then the 100% line is considered the base line.

Where the lowest value to be plotted on the Y scale is relatively high and a detailed scale is required to bring out the variations in all the data, starting the Y scale with zero introduces difficulties. For example, if we have a series of production figures over a number of years, ranging from 8,000,000 units to 10,000,000 units, then starting with a zero origin would have one of two undesirable consequences: either (1) the necessarily large intervals (say 4 million) on the Y scale would make us lose sight of the extent of fluctuation in the curve; or (2) a necessarily large graph to permit small intervals (say 100,000) would entail a waste of a large part of the graph, in addition to poor visual communication.

The solution is to break the Y scale. If the zero origin is shown, then the scale is broken by drawing a horizontal wavy line or a vertical wavy line between zero and the first unit on the Y scale, which in our illustration would be 8,000,000 units. If the zero line is not shown, one of several types of scale breaks may be used. Types of scale breaks are shown in Illustration 5.2. The main point in these is to make the reader aware that we have tampered with the base line.

Sometimes a break is presented in midscale, but it is not considered good practice since it distorts the visual impression.

Unfortunately, all too often we find arithmetic line diagrams which show neither zero line nor scale break. It must be emphasized that this omission results in a wrong visual impression, since (1) the lowest value plotted looks unduly close to zero and (2) the fluctuations assume different importance when compared with a true base line at zero and when compared with a line that happens to be the lowest line on the graph. An illustration of the failure to show a zero line or a scale break will be found on page 469 in Chapter 24.

2. *Equal Distances Mean Equal Magnitudes.* On an arithmetic scale, equal magnitudes must be represented by equal distances. This requirement holds for both the X -axis and the Y -axis, but

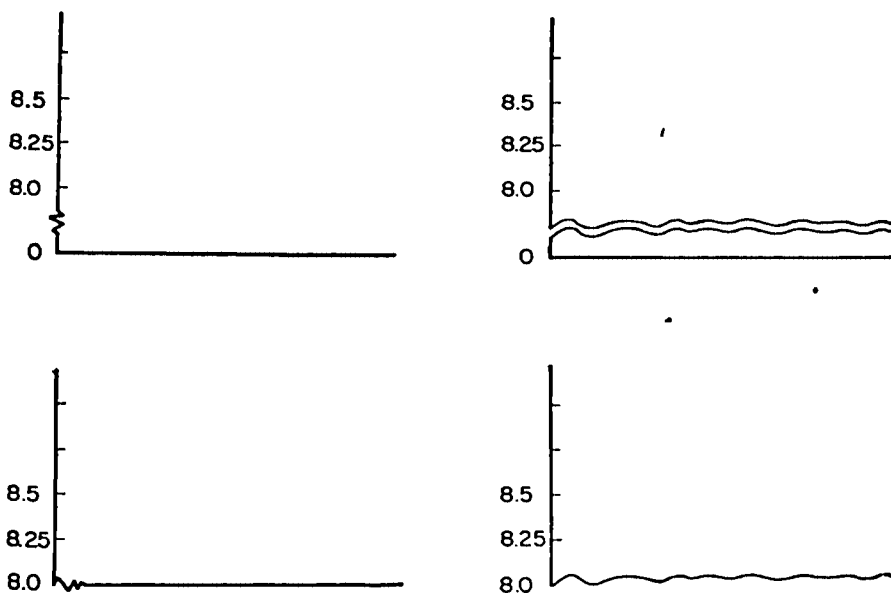


Illustration 5.2. Scale Breaks on Arithmetic Graphs.

for each separately. For example, on the Y -axis the distance from \$730,000 to \$740,000 must be equal to the distance on the same scale from \$1,010,000 to \$1,020,000. And on the X -axis the distance between 1950 and 1955, for example, must be equal to the distance between 1900 and 1905.

For that part of a scale which is broken, for obvious reasons, the above rule does not hold.

3. *One Vertical Scale Only.* It is preferable to use only one vertical scale on an arithmetic line graph. If two curves are shown, using for one curve a Y scale on the left and for the other a different Y scale on the right, a valid visual comparison of the two curves cannot be made.* We are comparing fluctuations in the two curves, and fluctuations depend on the scales. The fluctuations in a given series of data are of different amplitude on different scales.

If we use two different vertical scales on an arithmetic graph, one for each data series, we have no logical basis for comparisons.

* Of course, the *same* Y scale can be shown on both the left and the right of the graph.

Such use might involve an arbitrary assumption of a proportional relationship between the two series which is reflected in a proportional relationship in the two scales. Thus, if the scales are related as 1:3, then the assumption would necessarily be that the two series also are related as 1:3. Such has rarely been the case when two vertical scales have been used.

One solution is to transform the items of each of the two series into percents of some basic datum, and plot each series of percentages. The other is to use a semilogarithmic graph, the details of which will be discussed later in this chapter.

4. *Choice of Scale.* The choice of the scale is determined by an understanding of the data, and their fluctuations, as well as by the intention to give a clear picture of the data. Stretching the *X*-axis, for instance, will tend to de-emphasize fluctuations, while stretching the *Y*-axis will tend to overemphasize them.

5. *Plotting the Data.* Now that we understand the elements of an arithmetic line graph and the considerations that go into its construction, we are ready to discuss the actual plotting of the data.

For every *X* value there is a corresponding *Y* value. We wish to plot these matched values. Each pair of values is represented by a point which is plotted on the graph paper at that place which corresponds to the *X* value on the *X*-axis and the *Y* value on the *Y*-axis. Thus, if *X* is the year 1955 and *Y* is 10 tons, then the point is plotted as in Illustration 5.3. We repeat this procedure for each year in this illustration, and connect all the successive plotted points by straight lines. This succession of connecting lines gives us the curve.

The points may be plotted on the vertical grid lines or in the middle between vertical grid lines with the labeling and plotting in conformity. Where we are plotting monthly data for each year over a period of years, labeling only the years, we always label the spaces between the principal vertical lines, in order to avoid ambiguity. (See Illustration 5.4.) Analogously, if we were plotting daily data over a period of months, labeling

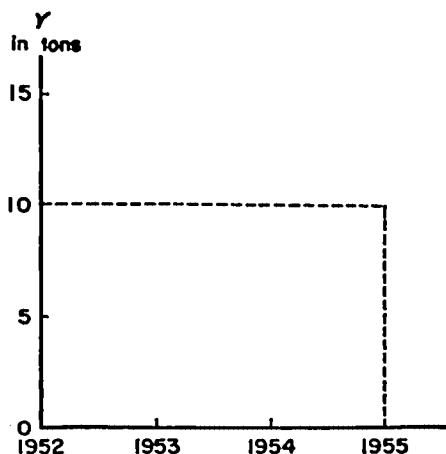


Illustration 5.3. How to Plot a Point on Arithmetic Graph Paper.

months, we would likewise label the spaces between the principal vertical lines.*

6. *Formal Requirements.* The formal requirements concerning title, headnote, footnote, and source note in a graph are analogous to those for a table.† Both the *X* scale and the *Y* scale are labeled. “Year” may be considered self-evident. In Illustration 5.4, the *Y* scale alone is labeled: “Sales in millions of dollars.” This label on a scale is called a scale caption. Scale captions on a graph take the place of the stub-head and the box-heads in a table. The scale caption for the *X*-axis is placed under the center of the horizontal axis. The scale caption for the *Y*-axis is placed at the top of the *Y* scale.

7. *Technical Aspects.* Ordinarily the curve and the axis lines are shown heavier than the rest of the graph. Grid lines guide the eye to the scale, but they are not indispensable. It is good

* In Illustration 5.4, the monthly data refer to the entire month, hence are plotted in the middle of the month spaces, that is, *between* the subsidiary vertical lines. If the data had referred to the first of each month, then they would have been plotted *on* a subsidiary vertical line.

† In printed material, however, it is customary to place all of these at the *bottom* of the graph.

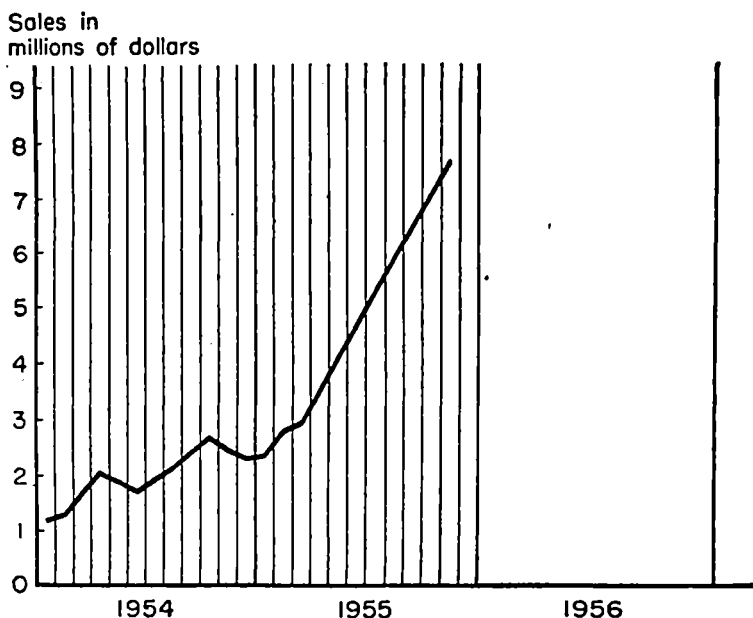


Illustration 5.4. Plotting Monthly Data for Each Year for a Number of Years.

practice to have as many grid lines as are necessary to guide the eye to the scale, but not so many as to compete visually with the curve.

If more than one curve is plotted on the same graph, then a different type of line must be used for each curve; for instance, full line, dotted line, dashed line. Each line should be labeled, or a legend should be added.

All lettering should be horizontal.

The grid may be drawn by hand, but the work of the statistician is made much easier if he uses printed graph paper.

A schematic diagram of an arithmetic line graph with all parts labeled will be found in Illustration 5.5, and an example of an arithmetic line graph, which compares two curves over time with one of them necessitating the use of the fourth quadrant as well as the first, is given in Chart 5.1 (page 50).

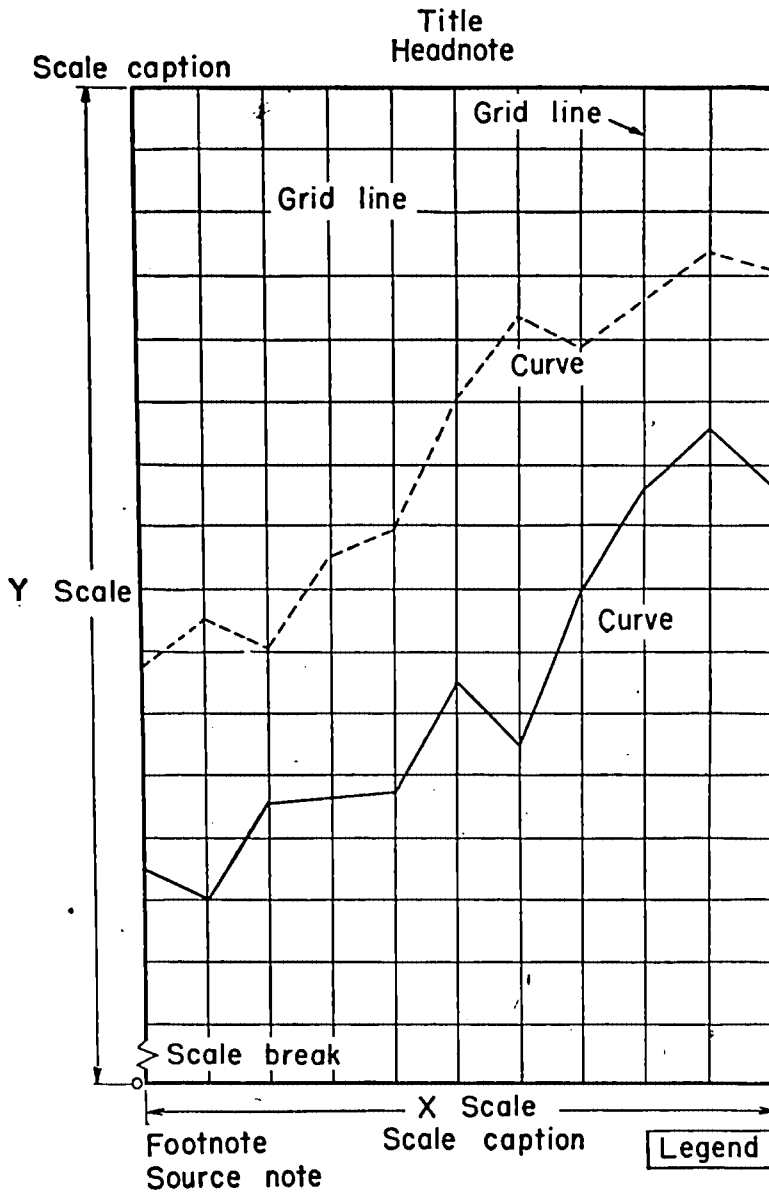


Illustration 5.5. Schematic Diagram of an Arithmetic Line Graph with All Parts Labeled.

Dollars per share

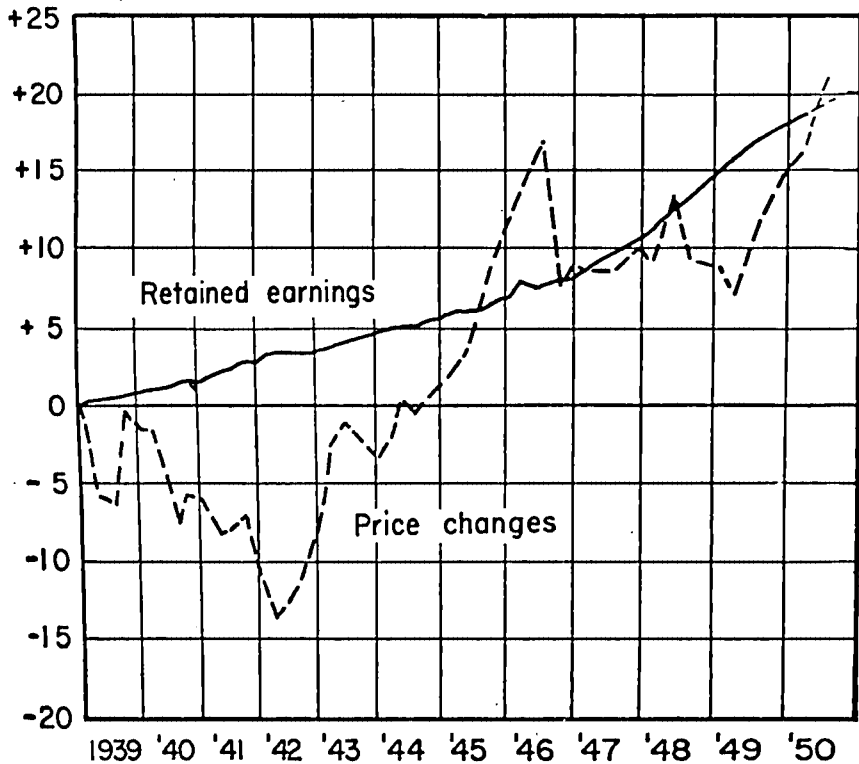


Chart 5.1. Corporate Retained Earnings and Changes in Stock Prices * (Cumulated Quarterly from 1938), 1939–1950.

Source: Moody's Investors Service; retained earnings per share computed by the Federal Reserve Bank of New York from Moody's data.

SEMILOGARITHMIC LINE GRAPHS

A semilogarithmic graph has one scale arithmetic and the other scale logarithmic.† Almost always the *Y* scale is logarithmic.

* Both retained earnings and price changes relate to the 125 industrial companies included in Moody's weighted average of 125 industrial stocks; second quarter of 1950 estimated by the Federal Reserve Bank of New York. Latest figure for stock price changes, June 22.

† A logarithm of a number (in the system commonly used) is the power to which 10 must be raised to produce that number. For example, $10^0 = 1$, $10^1 = 10$, $10^2 = 100$, $10^3 = 1000$. Therefore, the logarithm of 1 = 0, customarily written $\log 1 = 0$, $\log 10 = 1$, $\log 100 = 2$, $\log 1000 = 3$.

In passing, it should be mentioned that there is also a type of line graph that has both scales logarithmic.*

The Logarithmic Scale

In the arithmetic line graph, the scale is a series of consecutive units equally spaced. If, instead of such units themselves, we plot the logarithms of these successive units on an arithmetic grid, we have a logarithmic scale, as in Illustration 5.6. Since the logarithm of zero is minus infinity, this logarithm cannot be put on the scale. Consequently, no zero value can be shown on a logarithmic scale.

The distances between the points on the scale are the distances of the logarithms of the numbers from each other, and not the distances of the numbers themselves.

Construction of a Semilogarithmic Graph

Printed semilogarithmic or semilog graph paper is available ordinarily in from one cycle to four cycles. What does "cycle" mean here? It is the distance or interval along a logarithmic scale from a given value to ten times that value; for instance, from 1 to 10, from 2.8 to 28, from 0.36 to 3.6.† Such an interval is called a cycle or phase. We have as many cycles or phases on logarithmic paper as the number of times such a complete scale repeats itself. Thus, if we have to plot a series of figures from 3 to 300, we need 2-cycle paper—one cycle from 3 to 30 and another from 30 to 300. Illustrations 5.7, 5.8, and 5.9 show examples of one-cycle, two-cycle, and three-cycle semilogarithmic paper.

The number of grid lines indicating the subdivisions on the logarithmic scale differs from one manufacturer of semilog paper to another. For instance, one brand is made with ten subdivisions

* This type of line graph is sometimes called a double logarithmic graph, but it might better be called simply a logarithmic graph as distinguished from a semilogarithmic graph.

† Wherever feasible, the start of the semilog scale should be at some multiple of 5, to facilitate the plotting of the points.

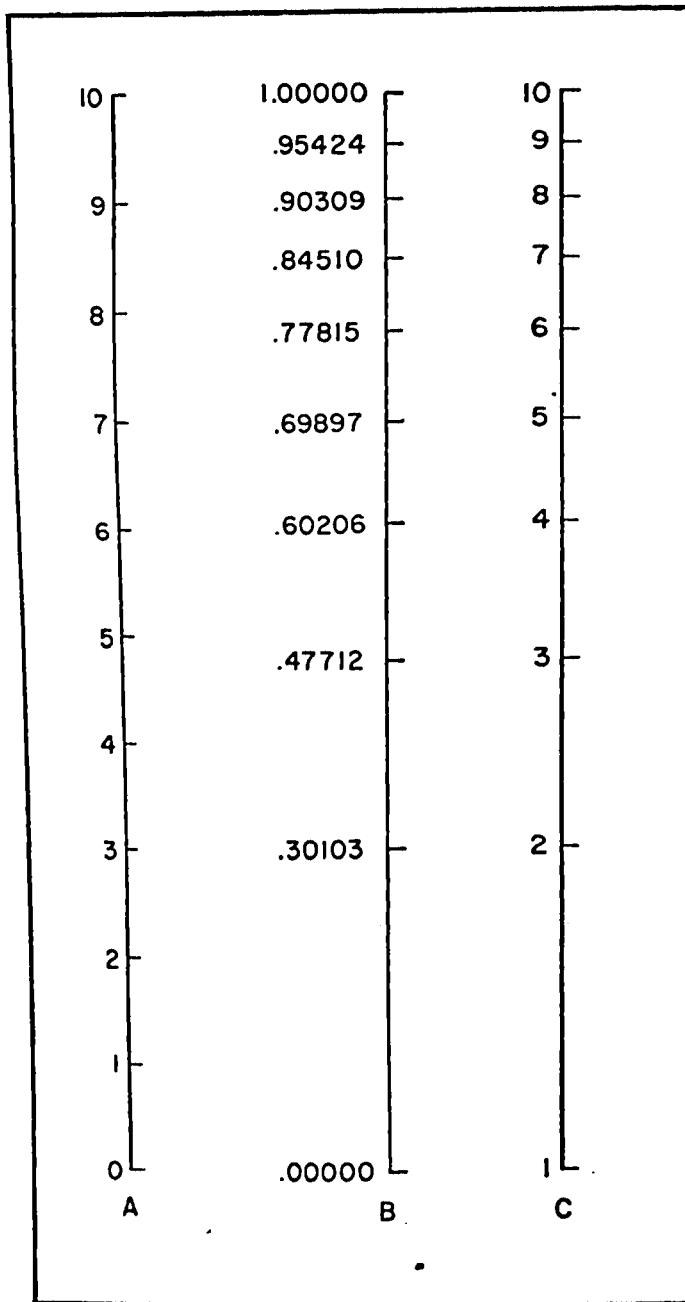


Illustration 5.6. Comparison of Arithmetic Scale and Logarithmic Scale: A, arithmetic scale; B, logarithms on arithmetic scale; C, logarithmic scale.

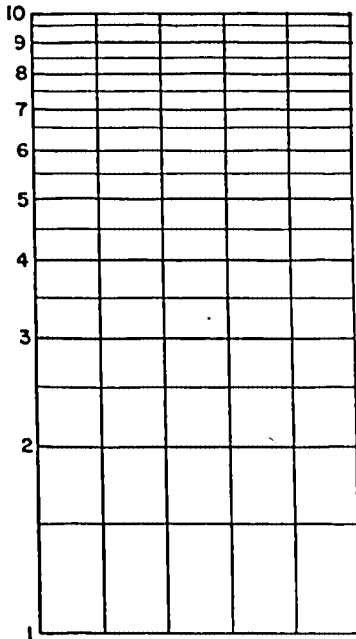


Illustration 5.7. One-cycle Semilog Graph Paper.

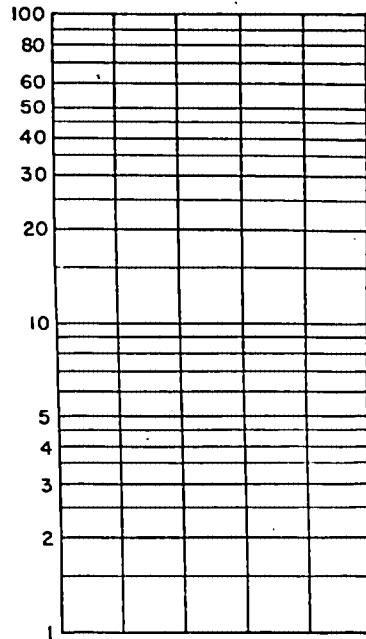


Illustration 5.8. Two-Cycle Semilog Graph Paper.

between 1 and 2 and another with twenty. This must be watched in laying out the scale.

Before plotting, we must label the scale with strict regard to the fact that the end of each cycle represents a value ten times that of its start. The heavy horizontal grid lines on printed semilog paper show the lowest value taken once, twice, three times, and so on; for instance, 1×100 , 2×100 , 3×100 , 4×100 , and so on, and then in the next cycle 1×1000 , 2×1000 , 3×1000 , 4×1000 , and so on.

To plot a given value on a logarithmic scale which falls *between* grid lines, we have to consider that values on this scale are spaced logarithmically, not arithmetically. Hence we must attempt the visual estimation of distances in terms of logarithms. Thus, if we wish to plot 3550, the point is not placed halfway between 3500 and 3600 but somewhat closer to 3600.

There is no need to present complete cycles if we do not use them. We can cut off the unused parts of the semilog paper.

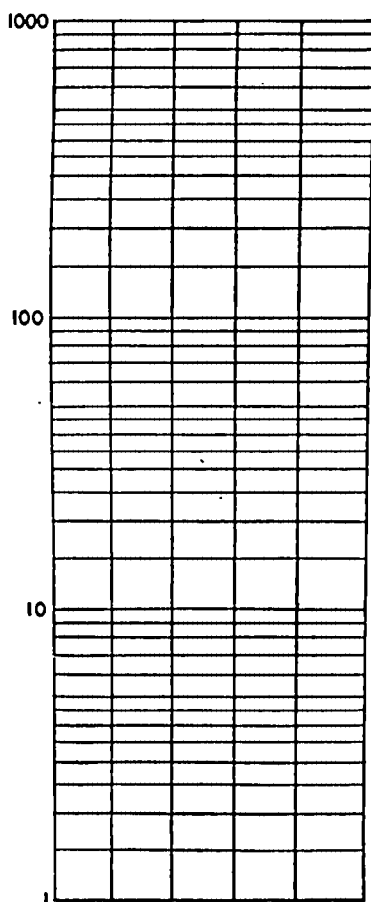


Illustration 5.9. Three-Cycle
Semilog Graph Paper.

Title, scale captions, source notes, and other formal requirements are the same for semilog line graphs as for arithmetic line graphs.

Characteristics of a Semilog Graph

Five characteristics of a semilog graph are guides for its use.

1. *Ratio Chart.* The semilog chart is a “ratio chart,” and is indeed thus designated. Equal distances represent equal ratios;

that is, the distance from 1 to 2 (a 100% increase) is equal to the distance from 2 to 4 (a 100% increase), from 4 to 8, and the like. (See Illustration 5.6.) On the arithmetic scale, on the other hand, the distance from 1 to 2 (a difference of 1) is equaled by the distance from 2 to 3 (a difference of 1), and so on.

2. *Rate of Change.* The semilog chart shows rates of change. If insurance in force in a given company increases from \$5 million to \$6 million in one year, and then \$1 million every year thereafter, plotting this change on a semilog chart makes clear the varying *rate of change*. Thus, the change from \$5 million to \$6 million means 20% increase, from \$6 million to \$7 million only $16\frac{2}{3}\%$ increase, and so on. If this increase of \$1 million each year were plotted on arithmetic paper, it would come out as a straight line, and successive increases of \$1 million every year would continue the straight line. But on a semilog chart, the increase every year over the preceding year would have to be 20% in order to get a straight line. That is, on semilog paper, a straight line will appear only if the *rate of change* remains the same for each time period. In short, semilog charts show *geometric* progression as a straight line, while arithmetic charts show *arithmetic* progression as a straight line.

3. *Comparisons.* The semilog chart is not appropriate for presenting and comparing absolute magnitudes or changes in absolute magnitude. For one thing, it cannot show negative magnitude or zero; there is no zero line. It shows slopes faithfully, and permits us to compare slopes usefully, whether of several curves or of several parts of one curve. If the series plotted in two curves are in different units, the fact is immaterial because it does not affect the slopes of the curves. Whether a curve is near or far from the bottom of the chart is also immaterial—the slope is the all-important consideration because it represents the relative developments and tendencies with which we are concerned when we use a semilog chart.

4. *Different Vertical Scales.* An increase of 15% shows the same slope whether it takes place in terms of hundreds or

millions, in tons or dollars. Hence, we may use different vertical scales for different curves in semilog graphs.

5. *Interpreting Slopes of Two Curves.* If two curves on a semilog chart plot as nearly parallel lines, whether increasing or decreasing, then the two series they represent change at about the same rate. Thus, if shoe production in one company plots parallel on a semilog chart to nationwide shoe production, then the company is keeping pace with national developments.

If two curves vary in slope on a semilog chart (or one curve at different times varies in slope), then greater steepness indicates change at a faster rate. For example, if rubber is being allocated by the government during wartime, and if the allocation of rubber to one tire manufacturer is going down at a faster rate than over-all civilian rubber allocation for tires, the differing rates would show up on a semilog graph as a steeper downward slope of the curve for this particular manufacturer as compared with the curve for all tire manufacturers. The given manufacturer may complain that he is not getting his share of rubber and support his claim with a comparison of these two curves.

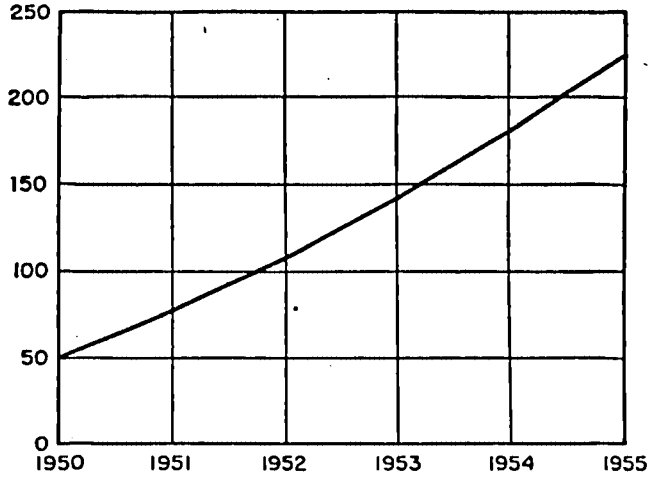
Uses of Semilog Graphs

1. *Rate of Growth.* If an investor were offered stock in Transatlantic Plastics Corporation, he might be impressed by the data in the first three columns below:

Year	Sales for Year	Increase over Sales for Preceding Year	
		Absolute	Percent
1950	\$ 50,000		
1951	75,000	\$25,000	50.0
1952	105,000	30,000	40.7
1953	140,000	35,000	33.3
1954	180,000	40,000	28.6
1955	225,000	45,000	25.0

A promoter might say, "Sales are growing faster every year!" But a statistician would say, "Not so! Sales are growing, but

Sales in
thousands of dollars



Sales in
thousands of dollars

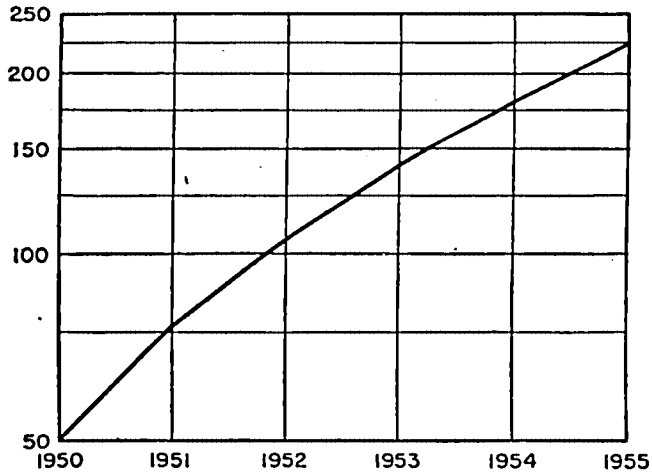


Chart 5.2. Sales of the Transatlantic Plastics Corporation, 1950-1955, on Arithmetic and Semilogarithmic Grids.

they are leveling off. Look at the rate of growth." The increase from 1950 to 1951 was \$25,000; that from 1954 to 1955 was \$45,000. But the earlier increase was 50% over the preceding

Millions of vehicles

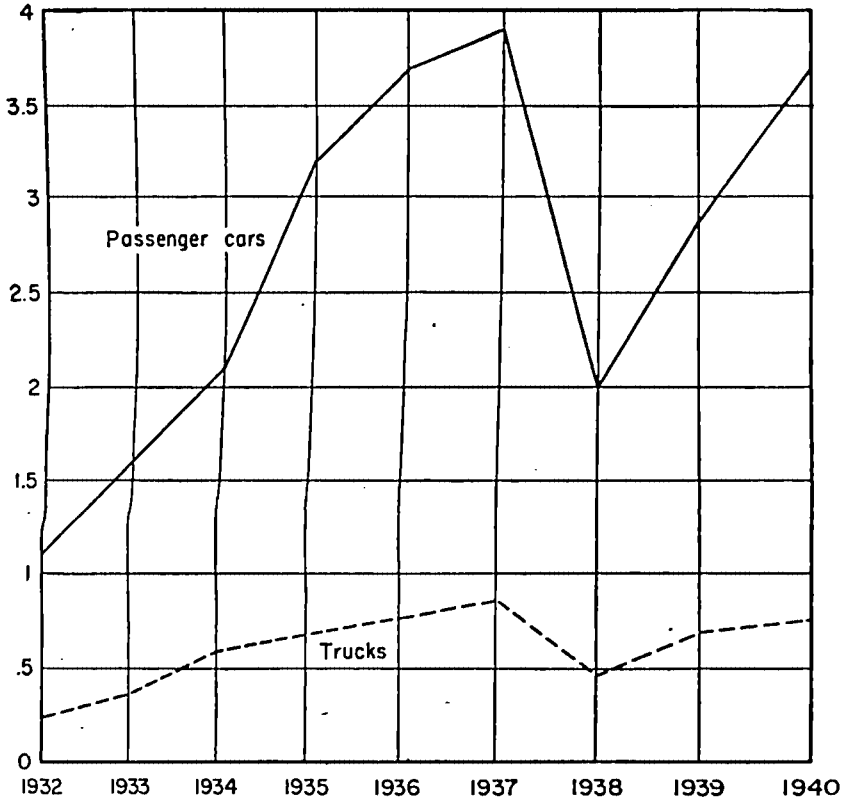


Chart 5.3. Comparison of Passenger-Car and Motor-Truck Production in the United States, 1932-1940 (on Arithmetic Grid).

Source: Automobile Manufacturers' Association, Detroit, Michigan, as reproduced in *The World Almanac*, 1950, p. 657.

year; the latest increase was 25%. The greatest absolute increase is the smallest percentage increase, a fact that appears instantly as the slope flattens on the semilog graph. But for this problem, plotting the data on the arithmetic line graph creates a slope as confusing as the absolute-increase figures. (See Chart 5.2.)

2. *Different Magnitudes.* Observe Chart 5.3, in which we compare truck production with passenger-car production for a number of years on one scale of an arithmetic grid. We find

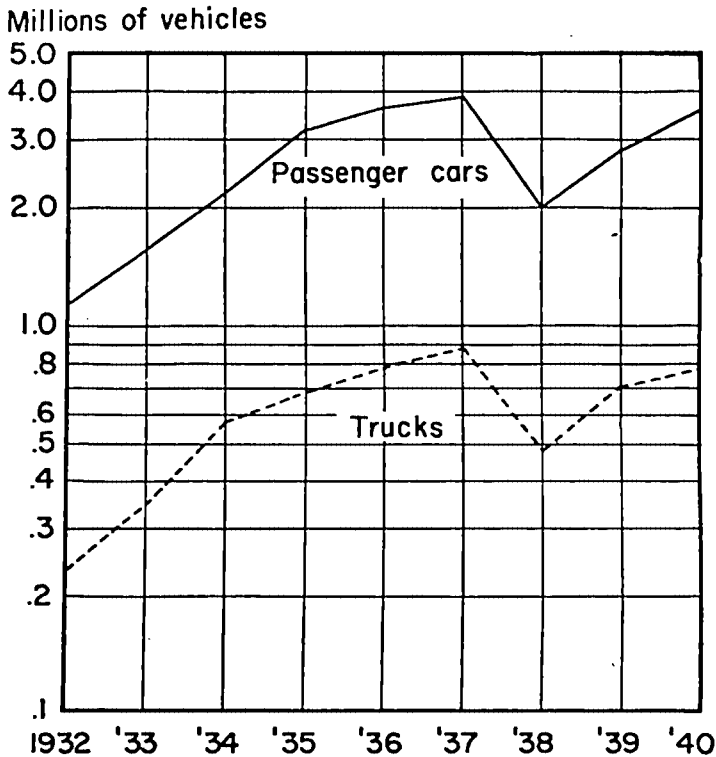


Chart 5.4. Comparison of Passenger-Car and Motor-Truck Production in the United States, 1932-1940 (on Semilogarithmic Grid).

Source: Same as Chart 5.3.

that the fluctuations of truck production appear to be relatively negligible; the curve of truck production appears to be creeping along the base line while that of passenger-car production fluctuates widely. The reason for the difference is that we are plotting absolute magnitudes on far different levels. An increase of 100,000 in truck production and the same increase in passenger-car production have far different significance. But on an arithmetic grid, 100,000 is always represented by the same distance. The semilog chart (Chart 5.4), however, permits us to overcome this visual distortion. We are able to represent an

increase of 100,000 in truck production as a *percentage* increase over the previous time period. Such changes, as shown on semilog paper, are not presented absolutely but rather in terms of their importance for truck production. Thus, we see from a semilog presentation that the movement of truck production actually parallels the movement of passenger-car production, because we are not presenting absolute numbers of vehicles produced, but rather the general tendency within each production series. (Compare Charts 5.3 and 5.4.)

3. *Unlike Units*. If we wish, for example, to compare nylon production in yards with that of nylon stockings in pairs, or to take another example, the amount of life insurance in force with the number of policies in force, presentation on a semilog grid is appropriate. This comparison cannot be graphed within the restriction to one vertical scale, which, we have seen, is inherent in the arithmetic grid. Thus on a semilog chart we achieve independence not only from absolute magnitudes but also from the units in which the data are expressed; therefore two different vertical scales may be used. Chart 5.5 illustrates the comparison of unlike units on a semilog chart.

4. *Comparison of Fluctuations*. If we wish to compare the development of one and the same series at different time periods, we can do so by means of a semilog chart. A steep rise or fall within the curve betokens large percentage change, whereas a flattening out means small percentage change. Thus, in Chart 5.4, considering now only the passenger-car production, we may compare the degree of fluctuation from 1935 to 1937 with that from 1938 to 1940. The first of these indicates a far more gradual development than the second.

5. *Study of Established Ratios*. Where there is an established ratio*—such as the “safe” ratio of 3:1 in some industries between current assets and current liabilities, or the corn-hog ratio of 11:1—the semilog chart permits us to ascertain easily and quickly whether the ratio has held throughout a given

* See Chapter 7.

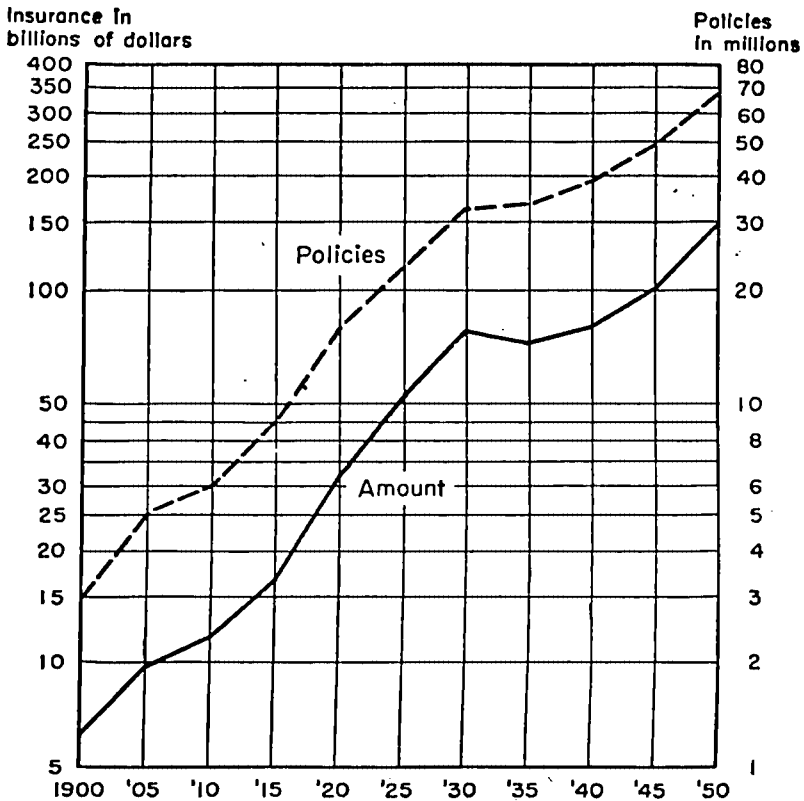


Chart 5.5. Ordinary Life Insurance Policies and Amount of Ordinary Life Insurance in Force in the United States, 1900-1950.

Source: *Life Insurance Fact Book*, 1955, p. 24; published by the Institute of Life Insurance, New York, N. Y.

time span. This can be done by arranging the Y scales according to the ratio in question. Let us illustrate this procedure by a hypothetical "safe" ratio 3:1 for current assets to current liabilities. The scale values for the assets on the semilog chart will be three times the scale values for the liabilities. Where the curves coincide, the ratio is 3:1. Where the curves do not coincide, the ratio is higher or lower than 3:1, depending on whether the assets curve is above or below the liabilities curve. This type of comparison is shown in Chart 5.6.

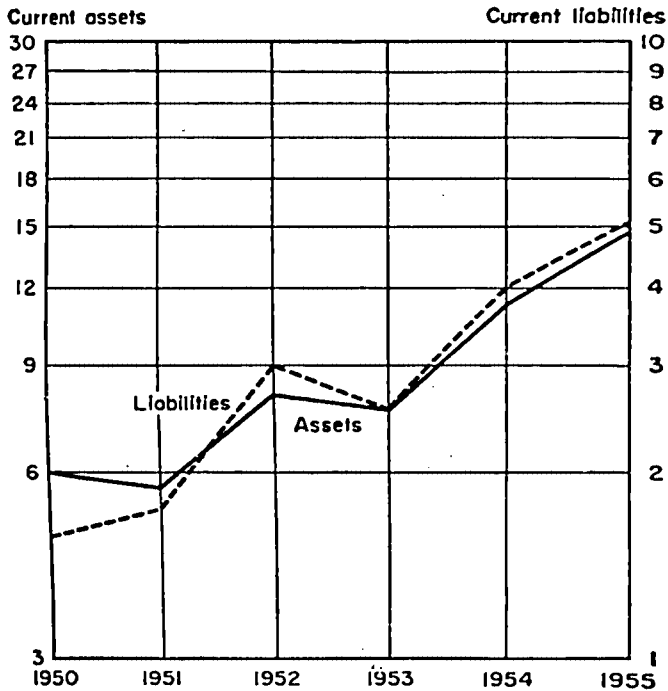


Chart 5.6. Comparison of Current Assets and Current Liabilities of the Television Corporation, Trenton, New Jersey, 1950-1955 (Hundreds of Thousands of Dollars).

6. *Reading Percentage Changes.* By means of an overlay scale, we can estimate percentage changes without computation. On a semilog chart, by placing the end of the overlay scale on the graph at any point of the curve, we can read off the percentage change from one period to another. Overlay scales to read percentage increases or decreases are shown in Illustration 5.10. The use of one cycle in such an overlay permits reading increases up to 900% and decreases up to 90%. If a cycle runs from 1 to 10 on semilog paper, then 10 represents an increase of 9 over 1, or 900% of 1. A decrease from 10 to 1 is 9 over 10, or 90% of 10. The increase of 9 here is compared with the small base of 1, whereas the decrease of 9 is compared with the large base of 10.

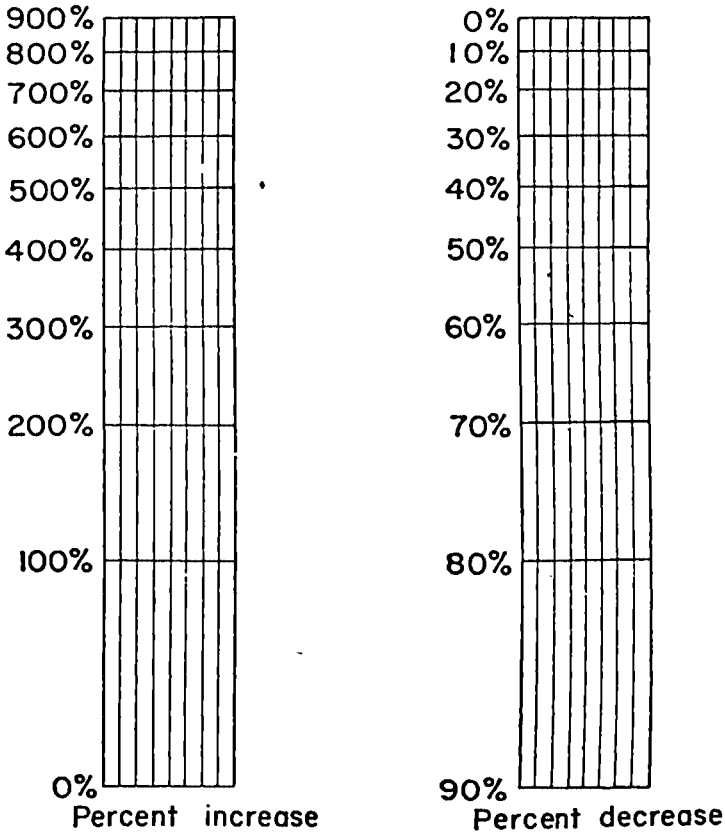


Illustration 5.10. Overlay Scales for Ascertaining Percentage Increase or Decrease on Semilog Graphs.

Limitations of the Semilog Graph

People familiar with semilog charts sometimes tend to overuse them. We must be sure that the basic characteristics of the semilog chart are fitting to our problem. For instance, if our problem requires us to relate and compare changes in absolute magnitude, the semilog chart must not be used. Ordinarily, the kind of data does not determine whether an arithmetic or semilog grid must be used, but what we mean to show and what we want to establish are the considerations that determine which one is to be used in a given problem.

Summary

1. There are two leading types of line graphs: arithmetic and semilogarithmic.
2. In constructing an arithmetic line graph, there are several rules and guides which refer to the zero origin, the meaning of equal distances, differing *Y* scales, the choice of scale, and plotting of the data.
3. Moreover, there are formal requirements as in tabular presentation, and there are technical aspects that must be considered in the construction of arithmetic line graphs.
4. The semilog graph has one axis (usually the horizontal) arithmetically scaled and one axis (usually the vertical) logarithmically scaled.
5. In a logarithmic scale, scale values are spaced according to the arithmetic-scale distances between their logarithms.
6. Logarithmic scales are divided into cycles of equal *linear* dimension, but each cycle ends at a magnitude ten times as high as that of its start, hence ten times as high as that of the preceding cycle's end—thus 1, 10, 100, 1000, etc.
7. On semilog paper, equal linear distances represent equal ratios. The semilog chart shows rates of change. There is no zero line and the slope is the all-important consideration. In comparing two or more curves, we compare their slopes.
8. Semilog graphs are useful in showing rates of growth, comparing series with widely different magnitudes and in unlike units, comparing fluctuations within the same series, checking established ratios, and reading percentage changes without computing.

CHAPTER

6

Statistical Presentation: Geometric Forms, Pictures, and Maps

GEOMETRIC FORMS

Bar Charts

A bar chart is a graph that consists of a number of rectangles (called bars) whose length varies with the magnitude represented, but which are of equal width. The bars may be arranged vertically or horizontally. The bar chart is used to show changes in magnitude over time or space, or in data that differ qualitatively.

Data classified geographically or qualitatively cannot be presented on a line graph, since they have only one numerical variable; a bar chart (or some other geometric form) must be used. A series of data over time, on the other hand, may be presented either in a line graph or in a bar chart. Accordingly, the arrangement or sequence of elements in geographic or qualitative bar charts depends on the purpose of the statistician.

A vertical bar chart showing changes in magnitude for selected states is illustrated by Chart 6.1. Chart 6.2 illustrates

Millions of Automobile Registrations

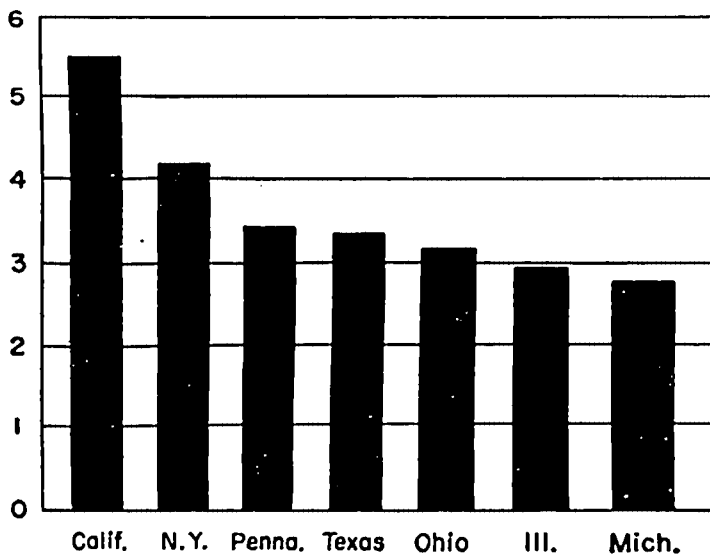


Chart 6.1. Automobile Registrations in Seven Selected States, 1953.

Source: United States Bureau of Public Roads, 1954.

a horizontal bar chart showing changes in magnitude over time.* As can be seen, bar charts are one-dimensional comparisons.

Simple bar charts consist of a number of *single* bars, each representing a magnitude of some variable. A group of *double* bars makes possible the comparison of magnitudes of one variable in two of its aspects or the comparison of magnitudes of two variables. The double bar usually shows no gap between its two parts. If we compare the automobile registrations in seven selected states, we get an assortment of single bars as in Chart 6.1. On the other hand, a comparison of manufacturing employment in six selected states for two different years is achieved through an assortment of double bars as in Chart 6.3.

We may also have triple-bar charts for comparing three

* It is to be noted that the omission of the years 1946, 1947, 1948, and 1949 in Chart 6.2 is correctly shown by a larger distance between the 1945 and 1950 bars than between any other two bars, though not proportionately so.

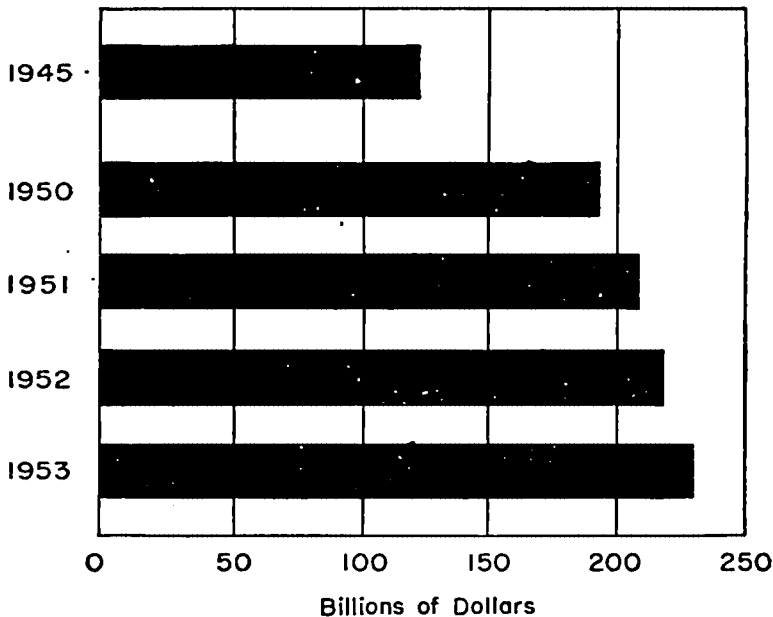


Chart 6.2. Consumer Expenditures in the United States, 1945–1953.

Source: United States Department of Commerce, Office of Business Economics, 1954.

aspects of a variable. With more than three bars, the bar chart becomes cumbersome and confusing.

In double- or triple-bar charts, the bars have to be distinguished visually by shading or coloring. A legend explaining the shades or colors must be included, as in Chart 6.3.

All rules and technical requirements for bar charts are analogous to those for arithmetic line graphs; for instance, zero line, equal distances for equal magnitudes, scales, scale captions, title, source note.

Area Diagrams

In area diagrams, the comparisons are two-dimensional. The geometric forms for area diagrams are chiefly squares, rectangles, or circles (see Illustration 6.1). Rectangles, which

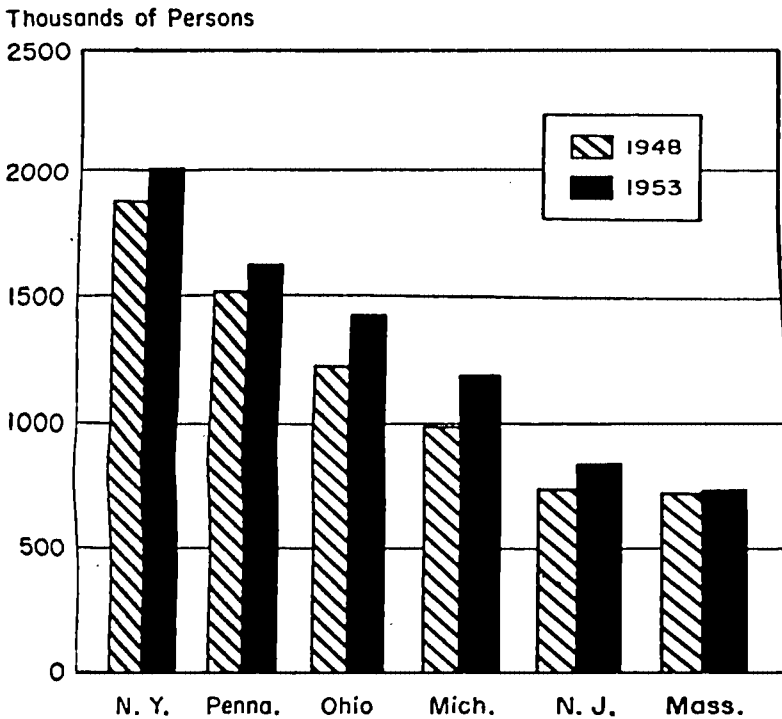


Chart 6.3. Manufacturing Employment in Six Selected States for 1948 and 1953.

Source: *Monthly Labor Review*, 1954, p. 694, and 1949, p. 458.

vary in width as well as length, are to be distinguished from bars which vary only in length. An ordinary observer can see that one area diagram is larger than another but few people if any can distinguish how much larger. Consequently, different magnitudes cannot be readily compared if presented as area diagrams. Therefore, two-dimensional diagrams are usually not a good form of statistical presentation.

Volume Diagrams

These may be cubes, spheres, cylinders, and other geometric forms involving three-dimensional comparisons. The limitations of visual comparison are even greater here than in area

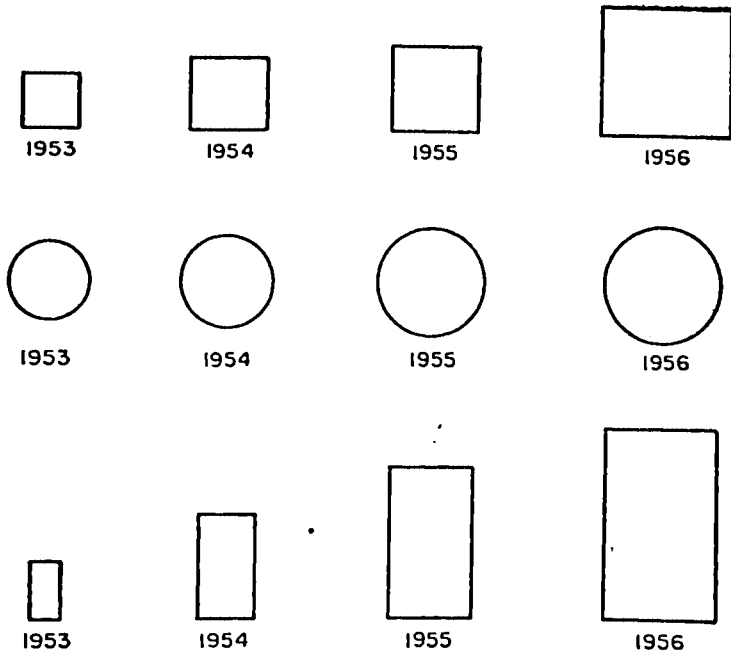


Illustration 6.1. Examples of Area Diagrams Showing Changes in Magnitude over Time.

diagrams. The use of volume diagrams for statistical presentation is not recommended.

The use of these volume diagrams for three-dimensional comparisons must be distinguished from the use of pictures. A cube, for example, when used as a symbol in a pictograph, is not presented for three-dimensional comparison.

PICTOGRAPHS

By means of pictures, we can show varying magnitudes. These pictographs are not abstract presentations such as lines or bars, but really depict the kind of data we are dealing with. If we are representing a number of farm owners, we show a symbol-picture of a farmer and each such complete symbol represents a stated number of farm owners; the total is ob-

tained by counting the number of symbols. If each picture stands for 1 thousand farm owners, and three such equal-sized identical pictures are presented for the year 1945, then there were 3 thousand farm owners during that year. If there are $3\frac{1}{2}$ thousand in 1955, then the additional $\frac{1}{2}$ thousand is represented by half a symbol. (See Illustration 6.2.)

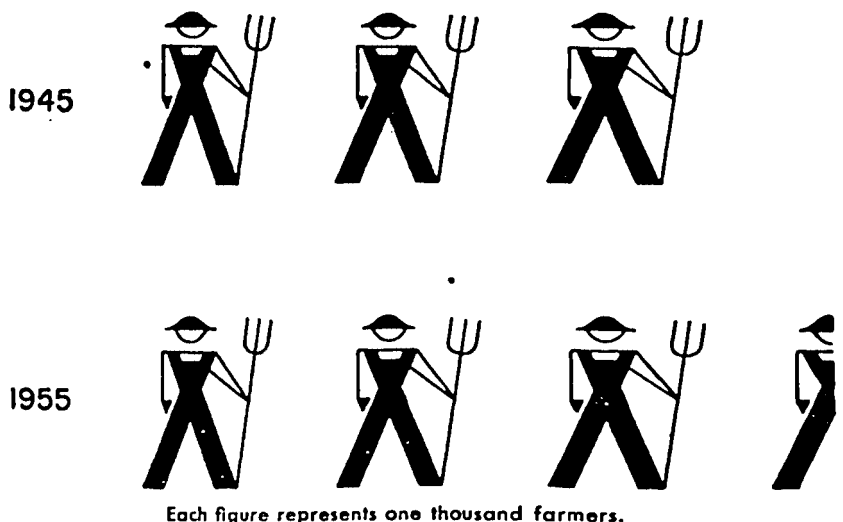


Illustration 6.2. Pictograph of the Number of Farm Owners, 1945 and 1955, in Blankandash County.

Pictographs are interpreted like bar charts, and in fact are modified types of bar charts. We interpret visually the length of the series of symbols in the same way as we interpret the length of a bar. In pictographs, however, we count, whereas in a bar chart we measure along a scale.

The essential rules* for constructing pictographs are: (1) Quantities are compared by means of the *number of pictures*, not the size of the pictures. Showing quantities by varying the size of the pictures would involve us in the difficulties, already mentioned, that apply to area or volume diagrams. (2) Symbols should be clear-cut and unambiguous. They should be

* See Rudolph Modley, *How to Use Pictorial Statistics*, Harper and Brothers, 1937, p. 15. Pioneer work in pictorial statistics was done by Otto Neurath.

self-explanatory. (3) Pictographs are not designed to illustrate one isolated item as, for instance, the national labor force in 1950. They are to be used for *comparison* of quantities, as, for example, the national labor force in each of five selected years.

This type of visual presentation is known as pictorial statistics. It is particularly useful in presenting statistics to the layman since it conveys information in a way he may find attractive, and is easy to comprehend (see Chart 6.4).

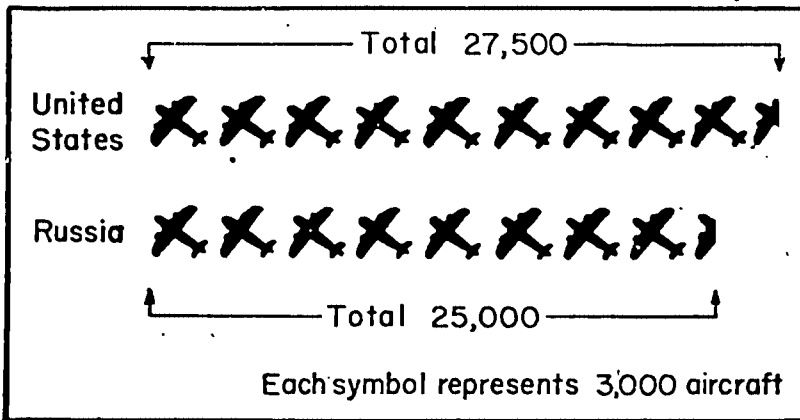


Chart 6.4. Comparison of Military Aircraft in Use by the United States and Russia, August 1950.

Source: Adapted from the *New York Times*, August 6, 1950, Section 4, p. E5.

MAPS

The purpose of statistical maps is to give quantitative information on a geographical basis, so as to facilitate comparisons by geographical areas. The quantities are usually shown in one of the following ways: (1) by shade or color; (2) by dots; (3) by placing bar charts, area diagrams, or pictographs in each geographical unit; (4) by placing the appropriate numerical figure in each geographical unit. These four types are illustrated in Charts 6.5, 6.6, 6.7, 6.8 respectively.

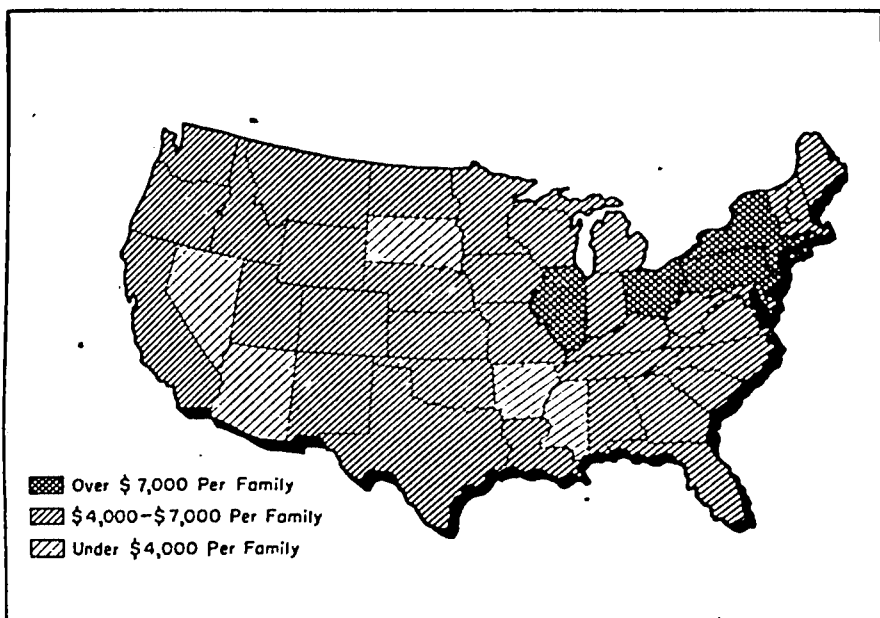


Chart 6.5. Life Insurance in Force per Family in the United States by State for 1954.

Source: *Life Insurance Fact Book*, 1955. Published by the Institute of Life Insurance, New York, N. Y., p. 10.

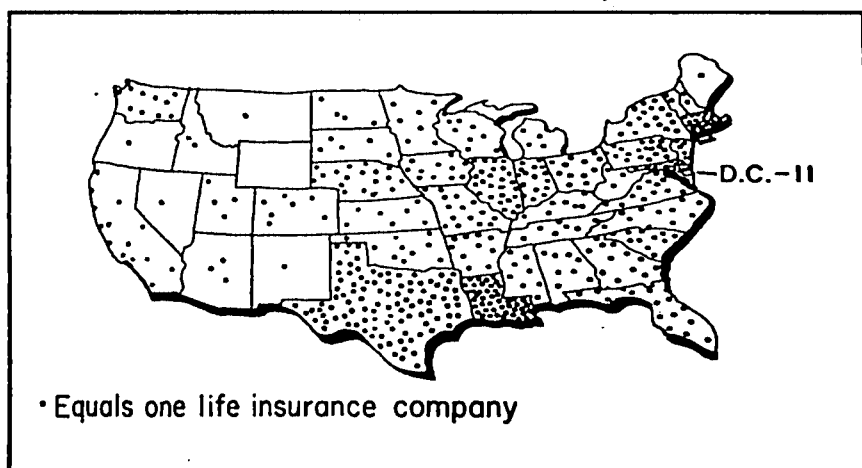


Chart 6.6. Number of Life Insurance Companies in the United States by States, January 1, 1949.

Source: *Life Insurance Fact Book*, 1949.

In constructing a statistical map it will usually be advantageous to use outline maps, which are available commercially or through governmental agencies.

Maps are useful in presenting comparisons of statistical data for different countries in the world, for different states in the United States, for different counties in a given state, and similar situations. From statistical maps the untrained observer quickly and easily gleans the pertinent statistical comparisons.

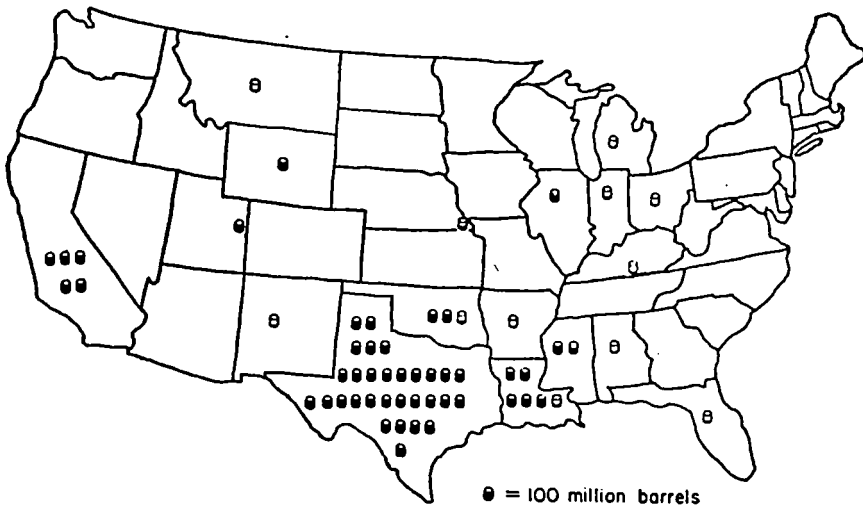


Chart 6.7. Net Additions to United States Petroleum Reserves by State, 1946-1949.

Source: *Fortune*, March 1950, p. 19.

COMBINATIONS OF DIFFERENT TYPES OF GRAPHS

We have pointed out that pictographs, bar charts, and other graphic representations can be combined with maps. Moreover, highly effective presentation can sometimes be achieved by some other combination of two types, for instance, bar chart and line diagram, as in Chart 6.9.

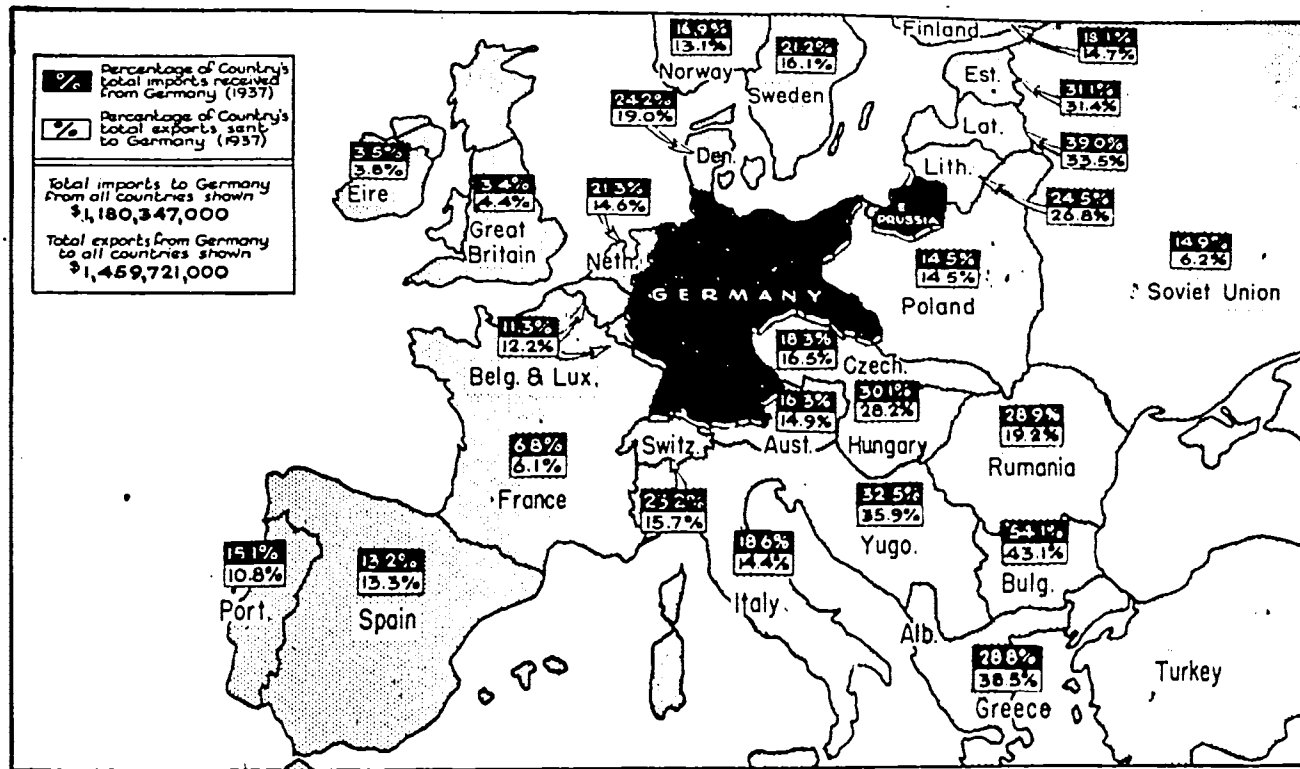


Chart 6.8. Germany's Prewar Foreign Trade with European Countries.

Source: *New York Times*.

Thousands of dozens

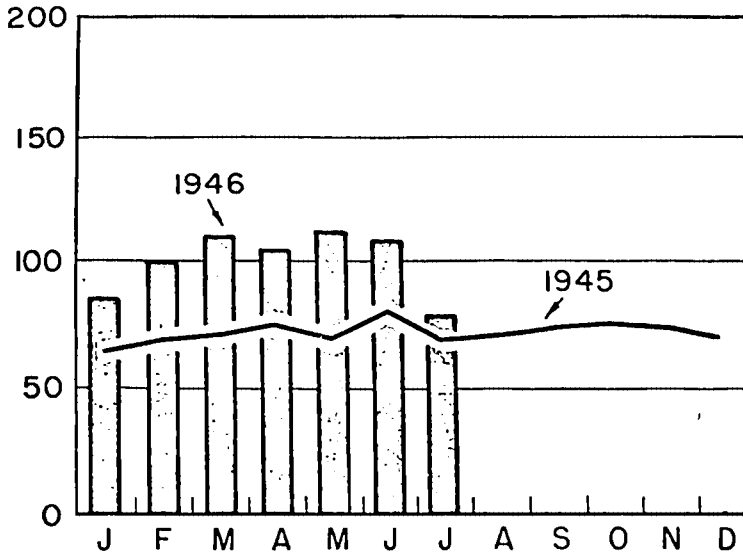


Chart 6.9. Average Weekly Production per Month of Knit Shirts in the United States, January 1, 1945, to July 31, 1946.

Source: *Facts for Industries Series*, Bureau of the Census, Industry Division, Apparel and Leather Unit.

COMPONENT-PART PRESENTATION

A problem that frequently arises in statistical presentation is how to communicate the breakdown of a total or series of totals. We wish to compare the changes over time that have taken place in the parts into which the total has been broken down and very often in the totals themselves. For example, a chain of variety stores has four branches. Component-part presentation makes possible the following comparisons:

1. Within any one year, a comparison of sales of each store with those of every other store of the chain.
2. Within any one year, a comparison of each store's sales with the total sales of the chain for that year.

3. From year to year, changes in the sales of each store compared with those of every other store.

4. From year to year, changes in the relative importance of each store in the total sales of the chain.

5. From year to year, changes in the sales of one particular store.

6. From year to year, changes in the total sales of the chain.

Such graphic comparison is illustrated in Chart 6.10 by component bar presentation on an absolute basis. These comparisons can also be presented by component pictographs, component line graphs, and pie diagrams.

The presentation may be in percentages rather than abso-

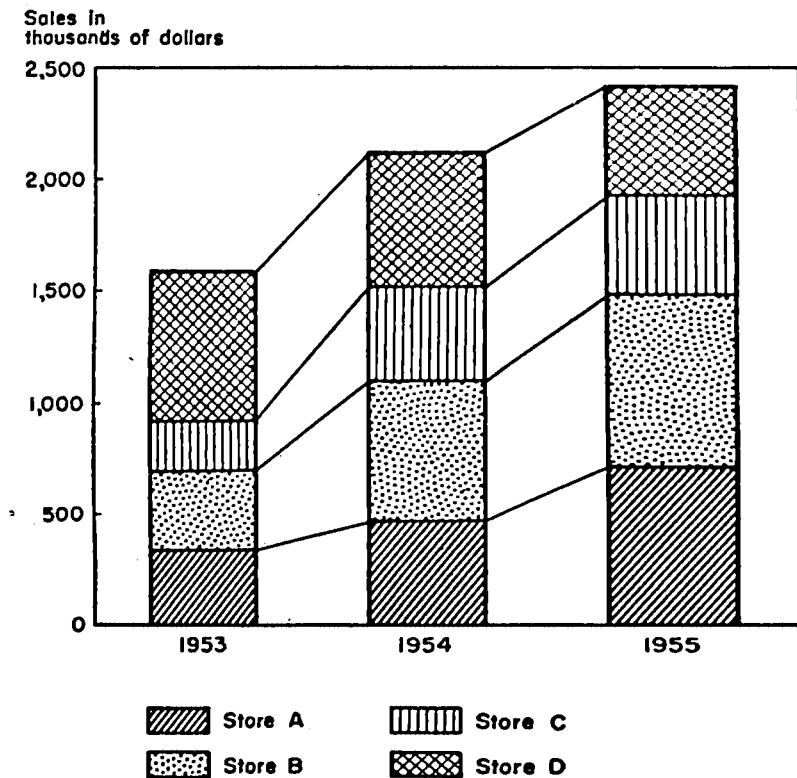


Chart 6.10. Annual Sales of the Amalgamated Minnesota Variety Stores, 1953-1955.

lute magnitudes. Total sales for each year then become 100% and the sales of each store in each year are expressed as percent of the total sales. The comparison of total sales from year to year is not possible if the presentation is in percentages since each year's total is always the same, namely 100%. Thus, on a percentage basis, the sixth point in the above list of comparisons does not apply.

Component Bar Chart

As can be seen from Chart 6.10, the length of the bar is broken up according to the size of the subdivisions. The component parts are differently shaded or colored, and a legend

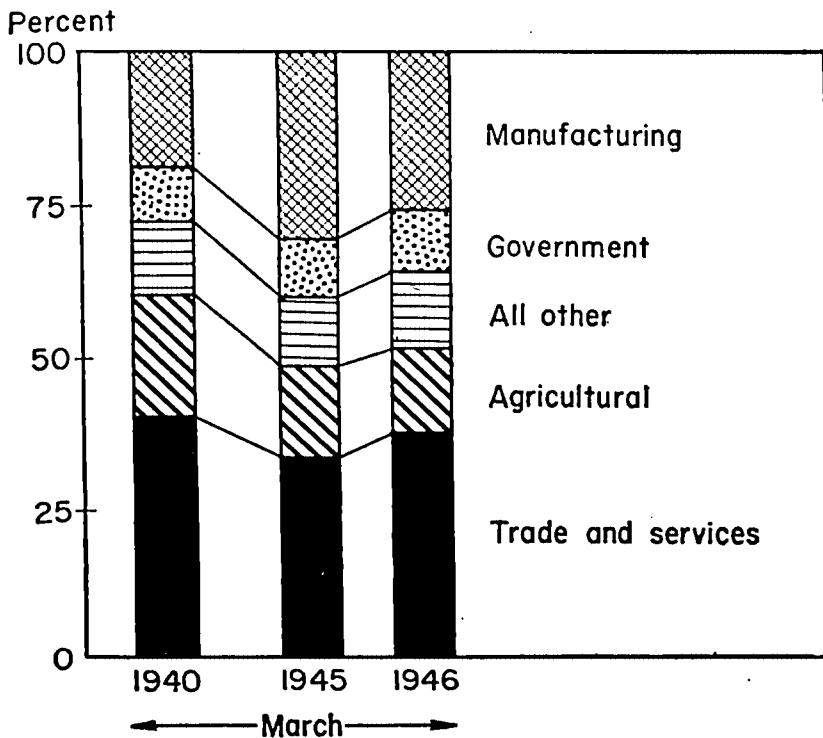


Chart 6.11. Percentage Distribution of Employed Civilians, by Industrial Groups, in the United States, 1940, 1945, and 1946.

Source: Adapted from *Survey of Current Business*.

may be added. In a series of component bar charts, it is customary to connect each subdivision of each bar with its counterpart in the adjoining bars, as is done in Chart 6.10.

If we have a series of percentage component bar charts, then all bars have the same length since the total is 100% in every case. Chart 6.11 illustrates this.

A component bar chart consisting of a single bar is illustrated in Chart 6.12.

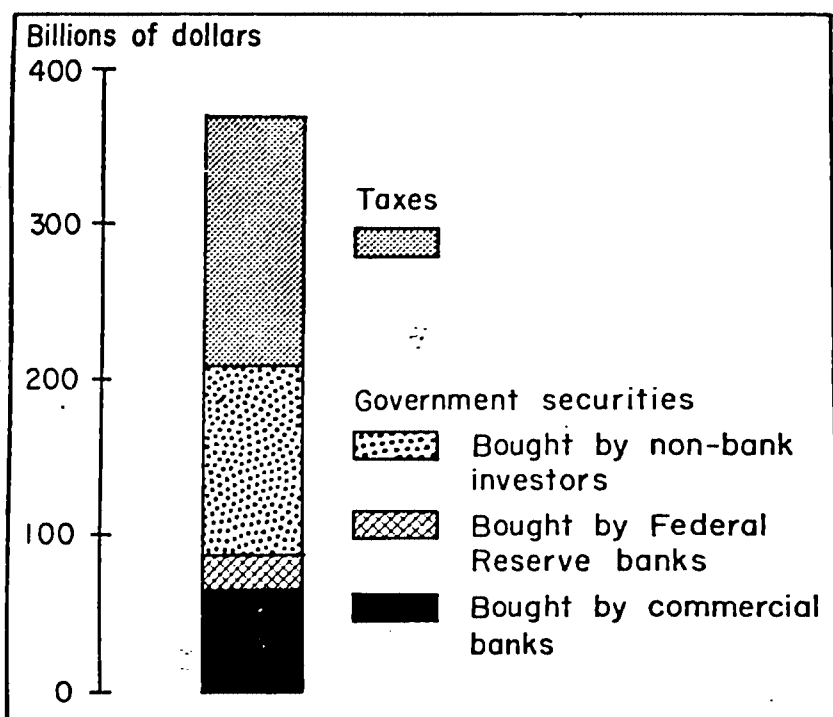


Chart 6.12. How the United States Government Financed World War II, July 1, 1941, to June 30, 1946.

Source: Adapted from *Our National Debt and the Banks*, No. 2 of National Debt Series by the Committee on Public Debt Policy, New York, p. 3.

Component Pictograph

The component pictograph has not been widely used. Here the component parts are shown by different symbols, but the

visual impression is comparable to that of a component bar chart. It is illustrated in Chart 6.13. The component pictograph may also be on a percentage basis.

Component Line Graph

In the series of component bar charts, we used connecting lines to join the same components in the different bars. A presentation similar to that of these connecting lines is the component line graph. This is illustrated in Chart 6.14 on an absolute basis and in Chart 6.15 on a percentage basis. In

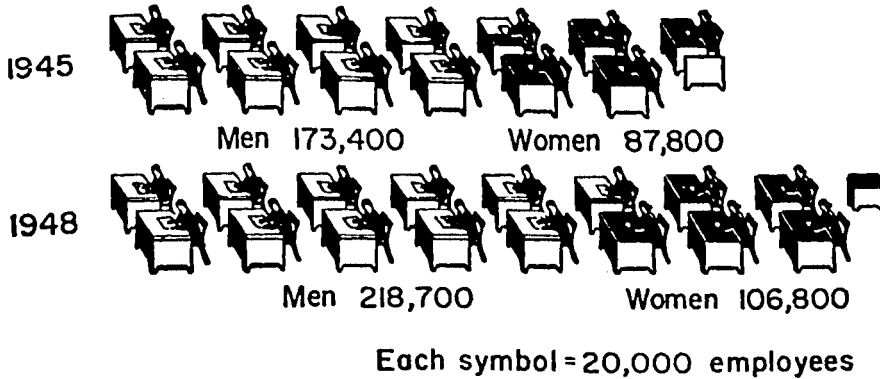


Chart 6.13. Total Employment in Life Insurance by Sex in the United States, 1945 and 1948.

Source: *Life Insurance Fact Book*, 1949, p. 72, published by the Institute of Life Insurance, New York.

Chart 6.14 the topmost curve shows the totals (as well as the last component plotted), and the other curves show the component parts. Where comparisons are to be made over a number of time periods, this type of graph may be used to advantage. Changes in the component parts are indicated by a narrowing or widening of the bands formed by the curves, and in fact this type of chart is sometimes called a band chart. These bands may usefully be distinguished from one another by different shades or colors.

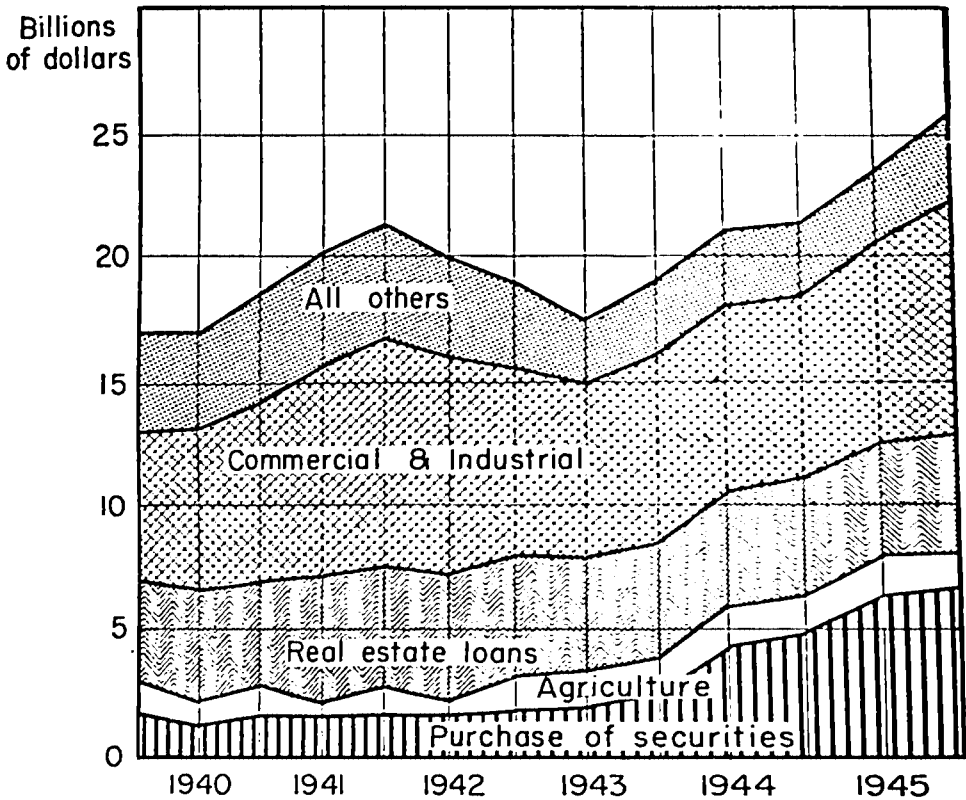


Chart 6.14. Loans of Insured Commercial Banks in the United States Classified by Use, 1940-1945.

Source: *Our National Debt and the Banks*, No. 2 of National Debt Series by the Committee on Public Debt Policy, New York.

Pie Diagram

A pie diagram is a circle broken down into component sectors. In comparisons, pie diagrams should be used on a percentage basis and not on an absolute basis, since a series of pie diagrams showing absolute figures would require that larger totals be represented by larger circles. Such presentation would involve us in the difficulties of two-dimensional comparisons (which we have already discussed in this chapter, under the heading of

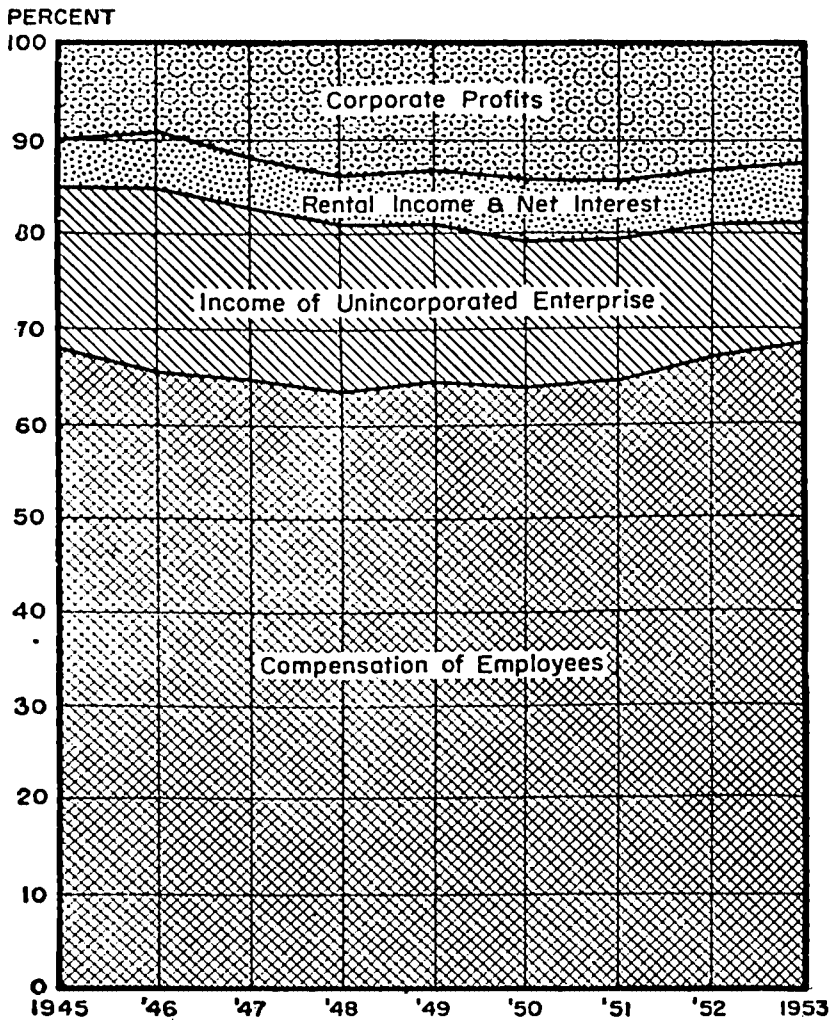


Chart 6.15. Percentage Distribution of National Income by Distributive Shares, 1945-1953.

Source: United States Department of Commerce, 1954.

“Area Diagrams”), whereas percentages can be presented by circles equal in size. Of course, this problem does not arise in the use of a single pie diagram.

In constructing pie diagrams, use of printed circles (shown in Illustration 6.3) with their circumferences divided in hun-

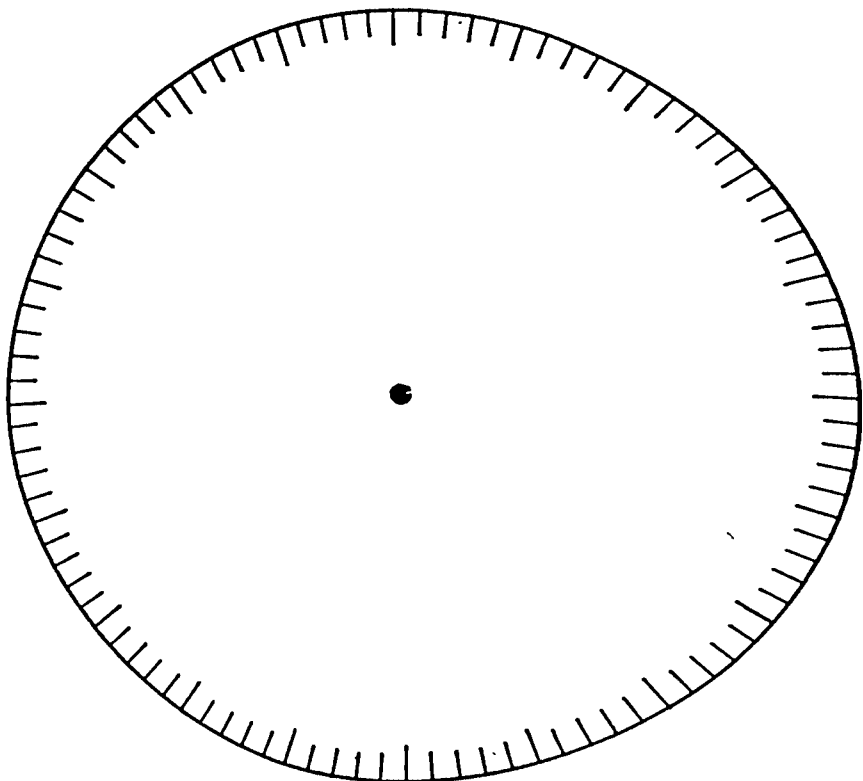


Illustration 6.3. Printed Circle for Construction of Pie Diagram.

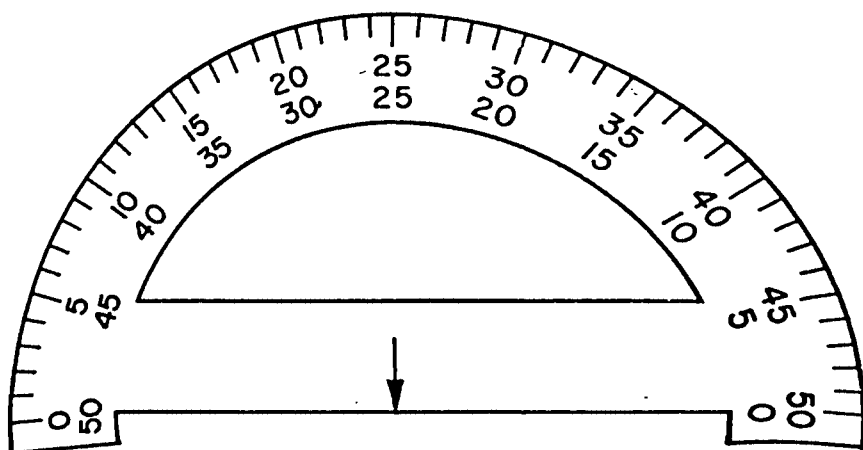


Illustration 6.4. Percentage Protractor.

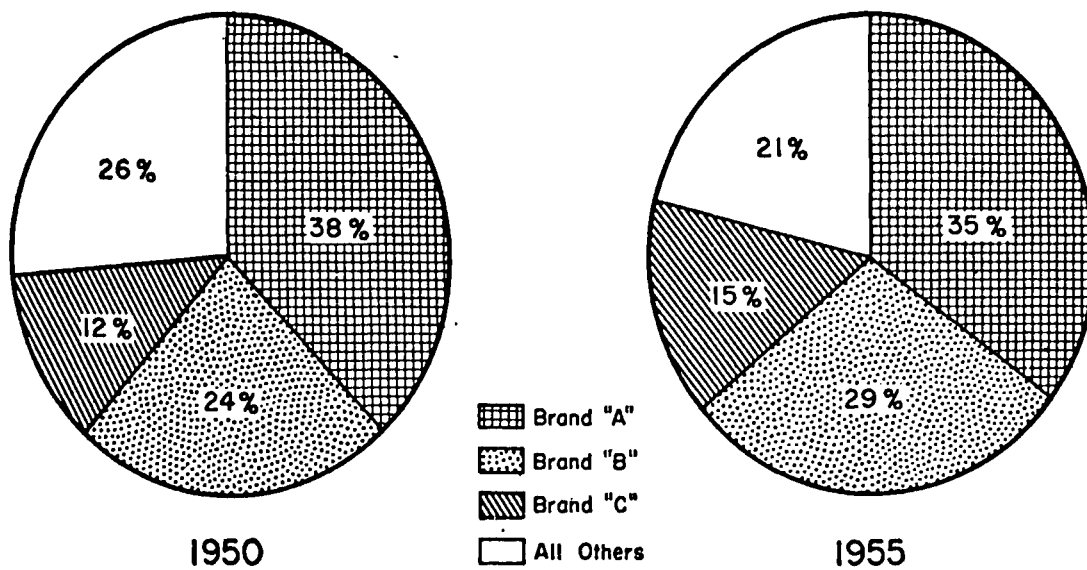


Chart 6.16. Market Shares of Various Brands of a Consumer Nondurable Good, 1950 and 1955.

Source: Confidential.

dredths will save much labor. A percentage protractor (shown in Illustration 6.4) may be used to lay off percentages of any circle.

The largest component sector of a pie diagram is usually placed beginning at the twelve o'clock position on the circle. Usually the other component sectors are placed in clockwise succession in descending order of magnitude, except for catch-all components like "All Others" and "Miscellaneous," which are shown last. Each component should be shaded or colored to contrast with adjacent sectors, when possible. The pie diagram is illustrated in Chart 6.16.

Choice of Component-part Diagram

Which type of component-part diagram is to be used depends on the data to be presented, the purpose of the presentation, and the characteristics of the various graphic forms discussed (lines, bars, pictures, pies).

FORMAL REQUIREMENTS

The formal requirements of graphic presentation for line graphs (discussed in Chapter 5) hold also for the types of graphic presentation discussed in this chapter, with adaptations necessitated by the peculiarities of each type of presentation.

MECHANICS OF GRAPHIC PRESENTATION

Lettering aids are available and should be used if the graph is to be reproduced. Wide use should be made of commercially printed graph paper in preference to hand construction. Pictographic symbols are also available commercially.

Neat graphic appearance can be achieved by either mounting or tracing the graph. In mounting we cut out the used part of the graph paper and paste it on white paper. The scales, title, and the like are then shown on the white paper. For

tracing, the graph is first made on graph paper; then a sheet of tracing paper or cloth is placed on top of it and all main lines traced.

The ultimate use to which the graph is to be put determines certain aspects of its construction. For instance, if photostats or one-color prints are planned, coloring should be avoided.

COMPARISON OF TABULAR AND GRAPHIC PRESENTATION

Statistical information ordinarily may be presented in both tabular and graphic forms. In deciding which form to use, we must keep in mind (1) that tables give precise figures whereas from graphs only approximate figures can be read; (2) that graphs give only a general impression but have eye appeal; (3) that tables usually require much closer reading and are more difficult to interpret; (4) that more information can be shown in one table than on one graph.

Often our aim will be to attract the interest and attention of the reader as well as to give precise information. Since precise information cannot be obtained from a graph, we then employ both tabular and graphic means of presentation.

The above considerations compare tabular and graphic forms from the standpoint of presentation to the consumer of statistics. But from the standpoint of the statistician, it must be noted that visualizing data in graphs may serve as a check on mathematical computation as well as a valuable guide toward analysis, and sometimes indeed as a tool of analysis. One critic of Adam Smith said that if Smith had only made a graph of certain facts he would not have misunderstood them.

Summary

1. The geometric forms used in statistical presentation are bar charts (one-dimensional comparisons), area diagrams

(two-dimensional comparisons), and volume diagrams (three-dimensional comparisons).

2. Differing magnitudes may also be compared by means of pictographs.

3. Statistical maps give quantitative information on a geographical basis.

4. Component-part presentation may be done through the component bar chart, the component pictograph, the component line graph, or the pie diagram.

5. Tabular and graphic presentation differ in their merits.

PART
3

Statistical Analysis

CHAPTER

7

Ratios

Meaning of Terms

A ratio is a comparison of one magnitude with another as a multiple or as a fraction. The main purpose of ratios is to simplify the numbers used in certain comparisons. If we compare the number of male workers with the number of female workers in the XYZ Corporation, we may express the comparison in absolute numbers as 355 male workers to 71 female workers. As a fraction this becomes $\frac{355}{71}$, or sometimes for typographical convenience 355/71. This may also be stated as a ratio of 355:71 or 5:1. This latter form for expressing a ratio is called a proportion.

"There are 284 more male workers than female workers" is *not* a statement of ratio.

It is often useful to express ratios with 100 as the base (or 10, or 1000, or still others). Thus we may prefer $\frac{500}{100}$ or 500:100 or 500/100 rather than $\frac{355}{71}$ or 355:71 or 355/71; all six forms are mathematically equal, but the ratios to 100 are perhaps more easily grasped and compared. A special and common example of such a ratio is the percentage. We could have said also that the number of XYZ's male employees is 500% of the number of its female employees. Ratios in such form are easily compared one with another. Thus we can say that at ABC Limited the number of male employees is 400% of the number

of female employees, whereas it is 500% in the XYZ Corporation.

Sometimes a ratio expressed as a percentage is called a *relative*. We shall meet in basic statistics terms such as "seasonal relative" and "price relative."

A ratio between two magnitudes usually shown over a period of time is called a *rate* if the magnitudes are qualitatively different even though expressed in the same units. Thus, an interest rate of 4% on a corporation's bonds means that for every \$100 of *principal* invested in these bonds an *interest* of \$4 a year is paid. A rate of speed of an automobile is a ratio of the number of miles traveled to the number of hours it took. We are all familiar with terms such as "birth rate" and "death rate." They signify that birth and death figures have been compared with population figures.

Types of Ratios

Ratios may be distinguished in terms of what is used as the base of comparison—that is, the denominator of the fraction.

1. We may compare a part to its whole. Thus, the sales in a selected department in a large retail store may be expressed in terms of the total sales of the entire store; we would say, for instance, that furniture sales are 43% of total sales. Whenever we have percentages of a whole, we may add them, and the total is 100%.

2. We may compare part to part within a whole. Thus, we compare the dollar volume of clothing sales with the dollar volume of furniture sales in one store, and we arrive at a statement such as "Clothing sales are 68% of furniture sales," or "Clothing sales are 68¢ for each dollar of furniture sales."

3. We may compare one whole to another whole. Thus, we compare the total sales in corporation A to the total sales in corporation B, and come to some such conclusion as "Total sales in corporation A are 80% of total sales in corporation B." The ratio may be an expression of the relation of one magnitude to a similar magnitude at the same time or place or at

different times or places. Thus, employment in Chicago may be compared for 1955 with 1954 as a base. Or employment in Chicago in 1955 may be compared with that in New York in 1955 as a base.

What we use as a base of comparison depends upon the purpose of our investigation. From the great variety of bases that may be used in establishing ratios, we have illustrated a few leading types.

Cautions Concerning Percentages

Ratios stated as percentages may give unsound impressions even though strictly correct mathematically. A conscientious statistician avoids such misuses of percentage, most of which are in five classes.

1. The base and the magnitude to be compared with it should not be too small. Thus, if there are only six executives in a corporation, and two are over the age of seventy, the statement that $33\frac{1}{3}\%$ of the executives are superannuated gives a false impression. The actual figures, 2 out of 6, constitute a better statement of the situation.

2. The magnitude to be measured against the base should not be too large (which may mean that the base is too small). Otherwise, we will arrive at a very high percentage figure which will not facilitate the comparison. To describe an increase in the resources of a bank during fifty years as 4000% does not appear to be a simplification and may even make comparison more difficult for the layman.

3. The magnitude to be compared should not be too small (which may mean that the base is too large). The statement that the number of workers in a given occupational group does not constitute more than $\frac{1}{80}\%$ of the population of one state and no more than $\frac{1}{80}\%$ of the population of another state makes comparison difficult. Here the statement of the absolute figures is probably preferable. To say that a water-supply bactericide will cause discomfort to 0.0002% of the population is less precise

and less clear than to say that about 1 person in 500,000 will have discomfort.

4. Shall changes in magnitude be expressed as ratios? This question cannot be answered absolutely, for the answer depends upon the problem being studied. In one case, percentages may reveal; in another, they may conceal. If a pencil sharpener has increased in price by 60¢ in three years (from \$1.00 to \$1.60), the statement that 60¢ is not such a great increase overlooks the fact that the increase was really 60%, which is sizable. Here the percentage increase reveals. On the other hand, if a new corporation, having shown very little profit the first year, reports an increase in profits of 3000% for the second year, this expression may conceal the fact that profits really increased by only \$500. Here, an accurate picture can be obtained only if the absolute figures are shown.

5. A comparison of percentage changes cannot be validly made without reference to their bases. If sales in a small outlet of a grocery chain increase by 40% from \$10,000, and the sales in a large supermarket of the same chain decrease by 40% from \$100,000, these two percentage changes, both of 40%, certainly do not cancel each other out. Total sales in both outlets combined have assuredly gone down, since the increase in the first outlet is only \$4000 while the decrease in the second outlet is \$40,000.

Some Important Ratios

Experience and statistical analysis have established the fact that certain ratios are important. Below are cited illustrations of some ratios in accounting, in agriculture, in personnel administration, and in management. These illustrations are, of course, not exhaustive, but do serve to show the importance of ratios as a statistical measure in economics and business.

1. Among accounting ratios, one that is well-known is the ratio of current assets to current liabilities. Thus, if the current assets of a corporation are \$3,000,000 and the current liabilities are \$1,000,000, then what is called the "current ratio" is

3.00. Certain standards have been set up for particular industries and businesses. These "safe" current ratios are considered guides for individual enterprises in these industries and businesses.

2. In agriculture, there are ratios such as the corn-hog ratio and the yield-per-acre ratio. The first means the dollar value of 100 pounds of live hogs compared with the dollar value of 1 bushel of corn. Based upon the amount of corn needed to raise a hog, a ratio of approximately 11:1 may be expected. When the ratio is above 11, it pays to raise hogs for corn is then used to greater advantage in raising hogs than in selling it on the open market. If the ratio is below 11, it pays to sell corn.

3. To compute labor turnover, the number of replacements required in one year is divided by the average labor force for that year. A high rate of labor turnover is undesirable since turnover entails expense for training new workers and discontinuity of personnel. This ratio may be subdivided into three different parts: the ratio of resignations to the average labor force, of discharges to the average labor force, and of lay-offs to the average labor force.

4. The ratio of sales value to costs gives an indication of operating efficiency in an enterprise. It is a comparison of dollar value of products sold and dollar production costs. This is an economic ratio which measures profits in a general way.

Summary

1. Ratios are comparisons of one magnitude with another as a fraction or as a multiple.

2. Ratios may take the form of fractions, proportions, percentages, or rates.

3. There are many types of ratio, which are to be distinguished by the base of comparison.

4. Percentages are widely used in economic and business statistics, but may be misleading if indiscriminately applied.

5. There are certain important ratios in current use in accounting, in agriculture, in personnel administration, in management, and in other parts of economics and business.

CHAPTER

8

The Frequency Distribution

Raw Data

The world, to be sure, is full of a number of things, in addition to cabbages and kings. We recognize as different things: incomes, workers, taxes, carloadings, ages, heights, births, deaths, retail sales, stock prices. But after we have distinguished things from each other and separated them according to their qualities, we are left with masses of items—each mass consisting of items having the same quality. But we are careful to see that each mass does consist of items of the same kind. If the thing we are concerned with is bank checking accounts, we may differentiate between checking accounts of individuals and checking accounts of corporations because, though both types are checking accounts, they may differ to such an extent as to constitute not a single quality but two separate qualities.

A mass of data in its original form is called *raw data*. A mass of data possessing a uniformity of quality with regard to the purpose of our investigation is known as *homogeneous data* or data of the same kind.

Each single item in the mass—as, for instance, one sales check, the wage of one worker, the price of one item—may be designated by various terms that are interchangeable one with

another. These terms are "value," "observation," "measure," "item," "case," "magnitude," "variate."

If we have obtained masses of items which are in numerical form, there is little that we can readily see except how many items we have that exhibit the quality we are interested in. If we knew all the wages paid to all the wage earners in a large industrial city, we would be staggered by this vast army of individual figures. Each figure is a wage, but what a variety of magnitudes! Thus, the raw data constitute an unorganized host of varying items. If we could get this large number of figures arranged so that they were in order of amount, we would still have the same number of items we started with, but we would know immediately which wage is highest and which lowest, what amount separates the highest from the lowest, and even begin to see where the largest part of these items appears to congregate and where gaps occur.

ARRAYS

The Simple Array

A mass of figures, which we have collected or which has been collected for us, when put into an orderly arrangement by magnitude (ascending or descending) is called an *array*.

A glance at the arrayed figures in Table 8.1 gives us the information we mentioned above. First of all, we now know that the lowest of these wages is \$30 and the highest is \$44. Second, the range between lowest and highest wage is \$14. Third, there is a concentration of wages between \$36 and \$39. Fourth, we notice a small gap near the beginning (no item of \$31) and a small gap at the end (no item of \$43).

With other data, it may be that in making an array we find that there is a concentration of numerical values among the low items or that there is a concentration of values among the high items. Moreover, the gaps may appear differently or may not appear at all.

TABLE 8.1. RAW DATA AND TWO ARRAYS OF WEEKLY WAGES OF 20 JUNIOR COPY TYPISTS IN NEW YORK CITY, APRIL 1949.

Raw Data				
\$34	\$39	\$36	\$40	\$30
41	36	44	38	33
39	37	42	35	39
30	38	36	37	32
Array in Ascending Order		Array in Descending Order		
\$30	\$37	\$44	\$37	
30	38	42	36	
32	38	41	36	
33	39	40	36	
34	39	39	35	
35	39	39	34	
36	40	39	33	
36	41	38	32	
36	42	38	30	
37	44	37	30	

Source: *Studies in Labor Statistics*, No. 2, National Industrial Conference Board, *Clerical Salary Survey of Rates Paid*, April 1949, pp. 10-11.

In dealing with a rather small number of items, the array can be very handy; but in dealing with hundreds or thousands, an array results in an unwieldy series of numbers. This unwieldiness requires that we condense the data.

The Frequency Array

If we find from the very making of the simple array that there is a repetition of values, it may prove rewarding to make what is called a *frequency array*. Such an array is made by listing once and consecutively all the values occurring in the series, and noting the number of times each value occurs. "Frequency" means the number of times a value appears in a series.

Table 8.2 shows a frequency array for the data in Table 8.1. This frequency array makes clear the concentration of items around certain values; we see quickly that ten of these twenty typists earn between \$36 and \$39 per week.

TABLE 8.2. FREQUENCY ARRAY OF DATA IN TABLE 8.1

<i>Wage</i>	<i>Frequency</i>
30	//
32	/
33	/
34	/
35	/
36	///
37	//
38	//
39	///
40	/
41	/
42	/
44	/
Total frequency = 20	

There are inherent limitations, however, in the simple array and the frequency array. First of all, neither one gives what may be called a synoptic view of the individual items; that is, we are still so close to the individual items in both cases—although less so in the frequency array—that we cannot see characteristics of the mass. In addition, either array may be too awkward and bulky. Since neither the simple array nor the frequency array gives us an idea of the characteristics of the group, we are unable to compare characteristics of different groups.

THE FREQUENCY DISTRIBUTION

Classes

If we take the data and establish classes—that is, ranges of values—we are able to make the series more compact and to clear the way for establishing the characteristics of the mass.

Every class is delimited by what are called *class limits*. The class limits are the lowest and the highest values that can be included in the class. These two boundaries of a class are known as the lower limit and the upper limit of a class. The lower

limit of a class is a value such that no lesser value can fall into that class; if the lower limit is \$30, let us say, then no value less than \$30 can fall into that class. The upper limit of a class is a value such that no higher value can fall into that class; if the upper limit is \$35, then any value greater than \$35 cannot fall into this class.

The width of a class is called the *class interval*. The method for establishing it is discussed later in this chapter.

Each class has a number of items that fall within the range of its interval; this number of items is called the *frequency* of that class. The mass of raw data has to be distributed over the classes set up. How do we go about this?

Tally Sheet and Entry Form

We may tally the data or we may use an entry form. Tallying consists of setting up classes and representing by a sloping or vertical stroke each item that falls in each class. When four such strokes have been made, a fifth horizontal stroke is drawn through them to represent the fifth item. In this way, bundles of fives are easily observable, and the process of totaling the frequencies in each class is expedited. An example of a tally sheet is shown in Illustration 8.1.

The entry form consists of a work sheet that has the classes

Tally Sheet of Distribution of Sales Checks in the Beta Store, Chicago, September 1, 1956.	
Sale in dollars	Number of Sales Checks
\$1.00 – 2.99	/// ③
3.00 – 4.99	/// // ⑦
5.00 – 6.99	/// /// ⑩

Illustration 8.1.

Entry Form for Sales Checks in the Beta Store, Chicago, Illinois,
September 1, 1956.

\$1.00-2.99	\$3.00-4.99	\$5.00-6.99	\$7.00-8.99	\$9.00-10.99	\$11.00-12.99	\$13.00-14.99
1.13	4.80	6.12	7.30	10.10	11.38	13.75
2.28	4.73	5.37	7.10	9.73	12.23	
1.90	3.20	5.62	8.36	9.20	12.40	
	4.12	6.34	7.80	10.50	11.00	(1)
(3)	3.00	6.67	8.98	10.42	11.85	
	4.00	5.03	8.75	9.69	12.01	
	4.83	6.45	7.10	10.99		
	(7)	5.90	8.32	9.27	(6)	
		6.25	7.47			
		5.00	7.65	(8)		
		(10)	8.00			
			8.80			
			8.80			
			7.15			
			7.50			
			(15)			

Illustration 8.2.

horizontally in sequence at the top, each with its lower and upper limit. Under each class is put the individual item—its exact value, not a stroke. The number of values that are found to fall under each class in the entry form is then counted and the count constitutes the frequency in that class. An example of an entry form is shown in Illustration 8.2.

It is obviously much easier to find class frequencies by tallying than through the entry form. But the entry form has two advantages over the tally sheet: (1) since we have the actual values of the items on an entry form, we can regroup into a new classification if the classes originally set up prove to be unsatisfactory, whereas from a tally sheet, we can only combine whole classes; (2) from an entry form we can check the accuracy of our entries, whereas in a tally sheet the items have lost their identity.

Neither the tally sheet nor the entry form is necessary if we have an array. To establish class frequencies from an array, we cut through the array at the points of the class limits.

The Frequency Table

Class frequencies, having been arrived at through a tally sheet, an entry form, or an array, can thereafter be assembled. The systematic presentation of the classes with the number of items falling within them is called a frequency table. An example of a frequency table is given in Table 8.3.

TABLE 8.3. FREQUENCY DISTRIBUTION OF SALES CHECKS IN THE BETA STORE, CHICAGO, SEPTEMBER 1, 1956.

<i>Sales in Dollars</i>	<i>Number of Sales Checks</i>
\$ 1.00– 2.99	3
3.00– 4.99	7
5.00– 6.99	10
7.00– 8.99	15
9.00–10.99	8
11.00–12.99	6
13.00–14.99	1
Total	<u>50</u>

Characteristics of Frequency Distributions

What have we done thus far? We have taken data that vary according to a measurable characteristic—in our example, the measurable characteristic is dollars—and put them into classes. Then we have ascertained the number of items falling within the limits of each class. This count of items constitutes the establishment of the frequency in each class. The number of items in all the classes combined is called the *total frequency* of the distribution. The data presented in the frequency table have been observed as of one point in time. A frequency distribution, therefore, is a snapshot of data, not a moving picture over time.

The Variable

Every frequency distribution involves the classification of a trait or quality that exhibits differences in magnitude; for example, prices, wages, number of employees, age, number of units produced or consumed, to mention a few of the almost unlimited traits or qualities which may form the basis of classification. The trait or quality which varies in amount or magnitude in a frequency distribution is called a variable.*

Some variables are capable of manifesting every conceivable fractional value within the range of possibilities; an example would be the weight of an industrial product. Such a variable is called *continuous*, as are the data involved. Other variables cannot manifest every conceivable fractional value but appear by limited gradations; for example, number of employees or number of machines in an industrial plant. Such a variable is called *discontinuous* or *discrete*, as are the data involved. In general, continuous data are arrived at through measuring, while discontinuous data are arrived at through counting.

In practice, certain types of discontinuous data are treated as though continuous if the gradations though limited are very

* The concept of the variable is not restricted, however, to frequency distributions.

small. That is, they are treated as if the series consisted of magnitudes that flow into one another. Such is the case with wage data expressed in dollars *and cents*. For practical convenience, data having very small discrete differences such as one cent are considered continuous. The difference of one cent is considered not a jump but a merging of values.

Mid Point

In classifying the sales checks on the entry form in Illustration 8.2, we found that 7 individual items fall in the class from \$3.00 to \$4.99. In a frequency distribution as in Table 8.3 (and in much practical work we are confronted with a frequency distribution, not the raw data), we see only that there are 7 items in the class from \$3.00 to \$4.99; these items have lost their individual identity.

Suppose we have at our disposal only the frequency distribution. Very often this is the form in which we get data. What, then, is the value of each of these 7 items, now anonymous? We are forced to make an assumption. Since the data fall between \$3.00 and \$4.99, we assume that they are spread evenly over this range (or are all located at the center of this range), and we take the value halfway between the lower and the upper limit. This value is called the mid point or mid value, and the assumption that makes it the representative of the class is called the mid-point assumption.

We obtain the mid point by adding the lower limit and the upper limit and dividing by two. This gives us as a rounded number a mid point of \$4.00 for the class \$3.00–\$4.99. Hence, \$4.00 is now taken as the value of each of the 7 items in this class.

Using the mid point as the value of each item in a class, instead of the original value of each item as in unorganized data or in an array, enables us to employ the grouped data (data in a frequency distribution) for computation. If we have only grouped data at our disposal, we cannot compute without making the mid-point assumption. And even if we have the

choice of computing from either grouped data or large numbers of ungrouped data (data unorganized or in an array), we will nearly always use the grouped data because computation is easier. The mid-point assumption makes such computation possible. We use grouped data especially where a great number of items is involved.

For the advantages of the mid-point assumption, just mentioned, we pay a price in loss of accuracy. If we total the original items placed in the class from \$3.00–\$4.99 in our example from Illustration 8.2, we get \$28.68. Dividing \$28.68 by 7 gives us \$4.10. Taking the mid value, \$4.00, as the representative value of the items in this class understates their values. Thus, we see that there is a difference between the average of the original items (\$4.10) and the average of the class limits (\$4.00). This discrepancy occurs here because the items are not evenly distributed over the class. But this type of discrepancy in some classes tends to be overcome in the frequency distribution as a whole by the fact that in other classes the mid-point assumption results in overstating the average of the original items.

PROBLEMS IN CONSTRUCTING A FREQUENCY DISTRIBUTION

Number of Classes

Faced with the raw data, or an array, we determine the number of classes in the light of the following conjoined considerations: (1) the number of items in the entire series; (2) the lowest and the highest values; (3) even distribution of items within the classes; (4) a regular sequence of frequencies; (5) the avoidance of an extremely small or an extremely large number of classes.

1. We need to know, first of all, how many items are to be classified. The number of items is one, but only one, determinant of the number of classes to be set up. Some statisticians lay great stress on the number of items as the way to determine a suitable

number of classes, but by itself the number of items is not sufficient for this purpose.

2. The range from the lowest to the highest value shows the compactness or the spread of the given number of items. If they are compact, a relatively small number of classes may suffice.

3. If the number of classes chosen were to lead to the establishment of classes with wide gaps between the items falling in each class, the class interval would be too large and the number of classes too small. For example, a class of \$30-\$39 would be too large for the items \$30, \$30, \$32, \$32, \$32, \$36, \$37, since the mid value would not be representative. In such a situation, two classes might be set up, one \$30-\$34 and one \$35-\$39.

4. A fundamental premise in the construction of frequency distributions is that there is an underlying basic pattern that the data assume in the mass, and that the larger the number of items in a series the closer they come to this pattern. Different types of data have different patterns. We assume that the data we have on a given trait or quality will approximate the basic pattern that is valid for such trait or quality.

If a given number of classes leads to irregularity in the *sequence of frequencies* as we move through the distribution, we probably have too many classes (which means too small an interval). If the frequencies are say 2, 5, 3, 12, 2, 6, 1, 4, 2, 1, a ragged distribution results that obscures the basic underlying pattern in the distribution of the data. We may approach this pattern here if we lump together two classes at a time, resulting in the sequence of frequencies 7, 15, 8, 5, 3.

5. If we have a very large number of classes, we tend to lose simplicity and smoothness; if we have a very small number of classes, we lose details by lumping too much information into one particular class.

It should be obvious at this point that the number of classes chosen for a given series determines the size of the class interval. For example, with a range from lowest (10) to highest (90)

of 80, should we decide upon 8 equal classes, then the class interval will be 10. Should we decide to set up only 4 equal classes, then the class interval will be 20.

Actual Class Limits

If we establish a class for items reported as from 25 to 49, the *nominal* class limits are 25 and 49, but the *actual* class limits may be different. The actual limits depend on whether the data have been rounded; and if rounded, how they have been rounded. Three possible alternatives are:

1. Data are not rounded, as in number of employees. Here the actual limit is the same as the nominal limit.

2. If, for instance, data are in pounds and have been rounded to the nearest pound, then the actual limits would be 24.5 and 49.5.

3. If, for instance, data are in years and are rounded to the last full year, as in age data, then the actual limits are 25 and "under 50."

The mid point in the first and second cases would be 37, while in the third case it would be 37.5.

Special Problems of Class Limits

1. In determining class limits, it is usually not possible, and it is not necessary, to have the lower limit of the lowest class coincide with the lowest value in the ungrouped data, nor to have the upper limit of the highest class coincide with the highest value in the ungrouped data.

2. In certain types of data, we find concentrations at certain values; for example, in cafeteria checks we will find many items at multiples of five, or in annual salaries of executives, many items at certain round figures, such as \$6500, \$7000, \$10,000, and the like. In such cases, the mid value (which is the representative of the items in the class) must coincide with these points of concentration. Consequently, class limits (which determine what the mid value is) have to be arranged so that this coincidence takes place. If we have concentrations at

thousands in salary—data such as \$5500, \$6000, \$6000, \$6000, \$6000, \$6000, \$6250, \$6400, then our class here should be \$5500 and under \$6500.

3. Classes have to be mutually exclusive; there should be no doubt about where to place or find an item. That is, in setting the lower and upper limits of each class there must be no overlapping. If wages are in dollars and cents, and the class interval is \$5, the upper limit of the first class beginning with \$5 should not be \$10 if the lower limit of the next class is \$10. Otherwise it is impossible to figure out where any individual item of \$10 falls.

On the other hand, care must be exercised not to exclude too much, as in classes written "over \$5 and under \$10," "over \$10 and under \$15." In this case, there is no class that includes \$10.

4. There are a number of different ways of writing class limits. Any way of doing this is acceptable if it involves no ambiguity. The following ways illustrate acceptable forms:

- (a) 1 to 4; 5 to 8; 9 to 12.
- (b) 3.0–5.9; 6.0–8.9; 9.0–11.9.
- (c) 10 and under 15; 15 and under 20; 20 and under 25.
- (d) 57—; 61—; 65—.

Form (a) above obviously cannot be used if data are in fractions. In that event, form (b) (or a similar type using common fractions) is appropriate. Form (c) is not to be used for discontinuous data; "and under" signifies so small a difference from the upper limit that it can be neglected and the upper limits are considered for practical purposes as 15, 20, and 25 in form (c). In form (d), we do not write the upper limit of a class but consider it to be hard by the lower limit of the next class; thus, in our instance (d) above, the class 57— has to be read as 57 and under 61.

Open-End Classes

An open-end class is a class lacking one limit. Thus, a lowest class lacking a lower limit and a highest class lacking an upper

limit are said to be open-end classes. A class of "under 50" and a class of "1000 and over" are open-ended. The presence of one or a small number of extreme values—values that create big gaps between themselves and the main body of the data—makes it impossible to run the classification-system consecutively so as to include them. To get these extreme values—low and high—into the distribution, we open-end the lowest and highest classes.

Wherever possible, open-end classes should be avoided. They create certain difficulties, which will become more apparent in later discussions, because (1) it is impossible to establish the mid value of an open-end class, (2) the lack of a mid value handicaps us in computation, and (3) the lack of one limit handicaps us in graphic presentation.

The Actual Class Interval

The size of the class interval in a frequency distribution is found as the distance from one lower limit to the next lower limit. It is not always possible to find the interval by subtracting the lower limit from the upper limit of a class; for example, if we have classes such as \$26–\$29, \$30–\$33, \$34–\$37, the interval is \$4 *and not* \$3.

Varying Class Intervals

When we group data, it may turn out that the size of the class interval chosen results, in certain parts of the distribution, in no frequencies appearing in certain classes and very few in others. In order to avoid gaps or an irregular flow of frequencies in these parts of the distribution, we vary the size of the class interval. Moreover, sometimes a uniform small interval would lead to an unwieldy frequency distribution, and give us unnecessary details for certain parts of the distribution; for example, in sales of a store, an interval of \$0.05 may be important below \$1 but relatively unimportant for sales items between \$1 and \$10.

Table 8.4 shows a distribution of daily wages in an industrial plant, with varying class intervals.

TABLE 8.4. FREQUENCY DISTRIBUTION OF DAILY WAGES IN AN INDUSTRIAL PLANT, MILWAUKEE, WISCONSIN, NOVEMBER 1, 1956.

<i>Daily Wage in Dollars</i>	<i>Number of Workers</i>	<i>Frequency Densities (\$2 Interval)</i>
4 and under 6	10	
6 and under 8	15	
8 and under 10	25	
10 and under 20	10	2
20 and under 30	6	1.2
Total	66	

What are the drawbacks in using a frequency distribution with varying class intervals? The chief drawback is that the classes cannot all be compared as to their frequencies, since there is no uniform interval. As a result of this, we cannot interpret the distribution, or present it graphically, or compute certain measures.

How do we overcome this? We estimate what the frequencies would be in each class if a uniform interval were used. Thus, for the frequency distribution in Table 8.4 we take \$2 as the uniform interval. Under the assumption that the 10 items in the class \$10 and under \$20, and the 6 items in the class \$20 and under \$30 are evenly distributed, we may break down the interval into 5 equal intervals of \$2 each. We then assign one-fifth of the ten items in the \$10 and under \$20 class to each of the five new \$2 intervals. This gives a frequency of 2 items for each of these five construed subclasses. Analogously, in the class \$20 and under \$30 we arrive at a frequency of 1.2 for each of the five classes formed from the larger class. The frequencies thus obtained are called *frequency densities* and on this basis we can compare classes as to their frequencies, interpret the distribution as a whole, present it graphically, and compute certain measures.

Instead of breaking down large intervals into small intervals (as we have done here), we may on occasion obtain uniformity

by combining small intervals into large intervals of equal size. In this case no assumption concerning frequencies is needed.

PERCENTAGE FREQUENCIES

An instructive way of comparing class frequencies within a single distribution, and necessary in comparing class frequencies

TABLE 8.5. DISTRIBUTION AND PERCENTAGE DISTRIBUTION OF WEEKLY WAGE RATES OF A SELECTED * GROUP OF JUNIOR AND SENIOR COPY TYPISTS IN NEW YORK CITY, APRIL 1949.

Weekly wage rate, dollars	Number of typists		Percent of total number of typists	
	Junior	Senior	Junior	Senior
26 and under 30	2		1.7	
30 and under 34	20	13	16.7	3.9
34 and under 38	32	38	26.6	11.4
38 and under 42	36	68	30.0	20.4
42 and under 46	20	76	16.7	22.8
46 and under 50	6	51	5.0	15.4
50 and under 54	3	41	2.5	12.3
54 and under 58	1	28	0.8	8.4
58 and under 62		10		3.0
62 and under 66		5		1.5
66 and under 70		2		.6
70 and under 74		1		.3
Total	120	333	100.0	100.0

Source: Adapted from *Studies in Labor Statistics*, No. 2, National Industrial Conference Board, *Clerical Salary Survey of Rates Paid*, April 1949, pp. 10-11.

* Selected for illustration only. Analyses of the complete data would of course yield results different from our selected group.

in two or more distributions based upon a very different number of total items, is to transform the absolute frequencies into relative frequencies. These class frequencies expressed relative

to the total frequency are called *percentage frequencies*. We arrive at percentage frequencies by dividing the frequencies in each class by the total frequency of the distribution, and express the frequency in each class as a percentage of the total. Percentage frequency distributions are illustrated in Table 8.5. As will be seen, two distributions with an appreciable difference in total frequency will not permit comparison. On a percentage basis comparison is made possible. (See Chart 9.4.)

CUMULATIVE FREQUENCIES

A factual study involves the level of wages of senior copy typists. One contention is that the vast majority are earning far less than \$42 a week. Another contention is that most are earning

TABLE 8.6. DISTRIBUTION OF WEEKLY WAGE RATES OF SENIOR COPY TYPISTS IN NEW YORK CITY, APRIL 1949.

<i>Weekly Wage Rate, Dollars</i>	<i>Number of Typists</i>
30 and under 34	13
34 and under 38	38
38 and under 42	68
42 and under 46	76
46 and under 50	51
50 and under 54	41
54 and under 58	28
58 and under 62	10
62 and under 66	5
66 and under 70	2
70 and under 74	1
Total	<u>333</u>

Source: Table 8.5.

more than \$50 a week. The frequency distribution of wages of senior copy typists is seen in Table 8.6. But this frequency table by itself cannot clarify this issue.

In order to clarify such an issue, we make what is called

TABLE 8.7. SENIOR COPY TYPISTS IN NEW YORK CITY EARNING SPECIFIED WEEKLY WAGE OR MORE, AND EARNING LESS THAN SPECIFIED WEEKLY WAGE, APRIL 1949.

Weekly wage rate, dollars	Number of typists earning	
	Indicated weekly wage or more	Less than indicated weekly wage
30	333	0
34	320	13
38	282	51
42	214	119
46	138	195
50	87	246
54	46	287
58	18	315
62	8	325
66	3	330
70	1	332
74	0	333

Source: Table 8.5.

a *cumulative frequency* distribution. There are two ways of cumulating frequencies—upward or downward.

A cumulative frequency distribution may be made on a “less than” basis or on an “or more” basis.

“Less than” cumulative or upward.

How many workers receive less than \$34.00?

Answer: 13.

How many workers receive less than \$38.00?

Answer: $13 + 38 = 51$.

“Or more” cumulative or downward.

How many workers receive \$30 or more?

Answer: All 333.

How many workers receive \$34 or more?

Answer: $333 - 13 = 320$.

A complete illustration of both "less than" and "or more" cumulative frequency distributions is found in Table 8.7.

Cumulative frequency distributions may be put on a percentage basis. Here the cumulative frequency in each class

TABLE 8.8. PERCENT OF SENIOR COPY TYPISTS IN NEW YORK CITY EARNING SPECIFIED WEEKLY WAGE OR MORE, AND EARNING LESS THAN SPECIFIED WEEKLY WAGE, APRIL 1949.

Weekly wage rate, dollars	Percent of typists earning	
	Indicated weekly wage or more	Less than indicated weekly wage
30	100.0	0
34	96.1	3.9
38	84.7	15.3
42	64.3	35.7
46	41.4	58.6
50	26.1	73.9
54	13.8	86.2
58	5.4	94.6
62	2.4	97.6
66	0.9	99.1
70	0.3	99.7
74	0	100.0

Source: Table 8.7.

is expressed as a percent of the total frequency of the distribution. Percentage cumulative frequency distributions are shown in Table 8.8.

We are now in a statistical position to resolve the dispute. From Table 8.8 we see that 35.7% of the workers fail to obtain a wage of \$42 a week and 26.1% are earning \$50 or more per week.

By use of a cumulative frequency distribution, questions such as the following may be answered: How many American families have an annual income of \$5000 or more? What percentage of industrial workers fail to earn \$1.25 per hour? How

many white-collar workers in a particular industry are over 50 years of age? How many machines wear out in less than five years?

Summary

1. In their original form the data on a problem are called raw data. These raw data must be homogeneous.
2. The simple array organizes the data without condensation. The frequency array is one step toward condensation.
3. Raw or arrayed data are ungrouped. To group data we must set up a frequency distribution. The mechanics of setting this up from raw data involve the use of a tally sheet or entry form.
4. The presentation of a frequency distribution in a frequency table shows the classes and numbers of items in each class (frequencies).
5. The distinction between a continuous and a discontinuous variable is often of importance in work with frequency distributions.
6. The mid point (or mid value) of a class is taken as representative of the items in the class. This involves the mid-point assumption.
7. The problems of constructing a frequency distribution concern the number of classes, the class limits, the class intervals, and sometimes open-end classes and varying class intervals.
8. For purposes of comparison, absolute frequency distributions may be transformed into percentage frequency distributions.
9. Cumulative frequency distributions tell us the number of items or the percentage of items that fail to attain or surpass a given value in the distribution.

CHAPTER

9

Types of Frequency Graphs

Frequency graphs include histograms, frequency polygons, and ogives. All of these require grouping of the data. Visual presentation of ungrouped data, essentially a picture of the frequency array, may be accomplished through an array chart.

Array Charts

The upper part of Chart 9.1 is an array chart of the data concerning junior copy typists in Table 8.1. Compare it with the frequency array in Table 8.2. The lower part of Chart 9.1 is an array chart for the same number of senior copy typists.

The characteristics of data brought out by the array and the frequency array can be communicated effectively on an array chart. It shows the concentration of individual values, the spread of the series, whether the spread shows gaps or a uniform flow, and the location of items which are extreme. The array chart may be especially useful in comparing the above-mentioned characteristics in two or more series, as is shown in Chart 9.1.

GRAPHIC PRESENTATION OF FREQUENCY DISTRIBUTIONS

Thus far, we have presented the frequency distribution in tabular form. It is possible and rewarding to present the fre-

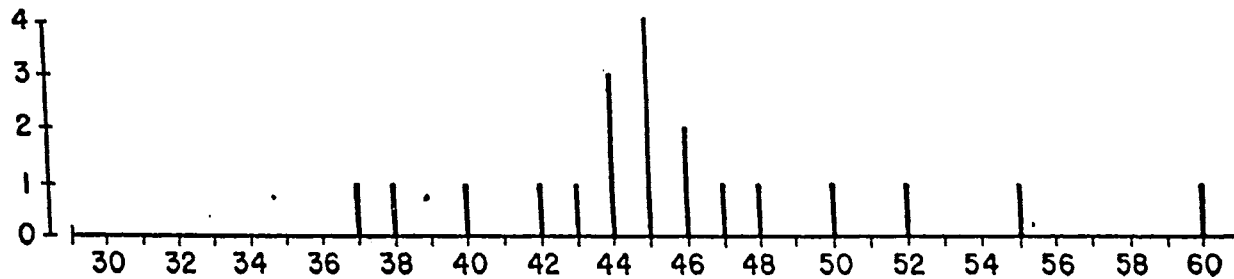
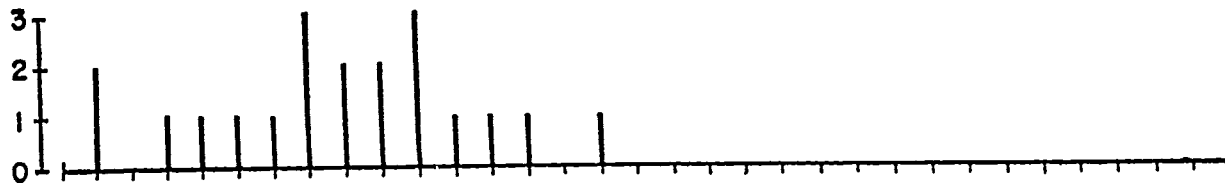


Chart 9.1. Two Array Charts.

quency distribution graphically as well. Thus the advantages of graphic presentation, already discussed, are obtained. The frequency-distribution graph has more eye appeal than the frequency table. It quickly calls attention to high spots and low spots in the distribution, and offers a vivid picture of characteristics of given frequency distributions.

Visualizing the distribution of the data is also of importance for planning the analysis. Frequency graphs make it easy to answer such questions as: What is the shape of the distribution? Is there just one concentration? Is there a pattern?

On a graph, we can present a frequency distribution (1) in absolute numbers or in percentage form, or (2) in a cumulative form. In absolute or percentage form we use what are called a histogram and a frequency polygon. For a cumulative distribution we use what is called an ogive.*

Histogram

The term histogram is formed from two Greek words, one meaning "something set upright" and the other meaning "drawing." In statistics, a histogram is a graph that represents the class frequencies in a frequency distribution by vertical rectangles.

On the *X*-axis we place the classes. On the *Y*-axis we show the frequencies which depend on the classes and therefore constitute, as it were, the dependent aspect.

The scale on the *X*-axis expresses class intervals, and each class is represented by a distance along the scale that is proportionate to its class interval. These distances are the widths of the vertical rectangles, and if all the class intervals are equal all the rectangles will be of the same width; if the class intervals vary, so will the widths of the rectangles. The scale on the *Y*-axis expresses frequencies and each class frequency establishes the height of its rectangle. Thus we get a series of rectangles, each having a class-interval distance as its width and a frequency distance as its height. The combination of these four-sided

* "Ogive" is pronounced ō' jiv.

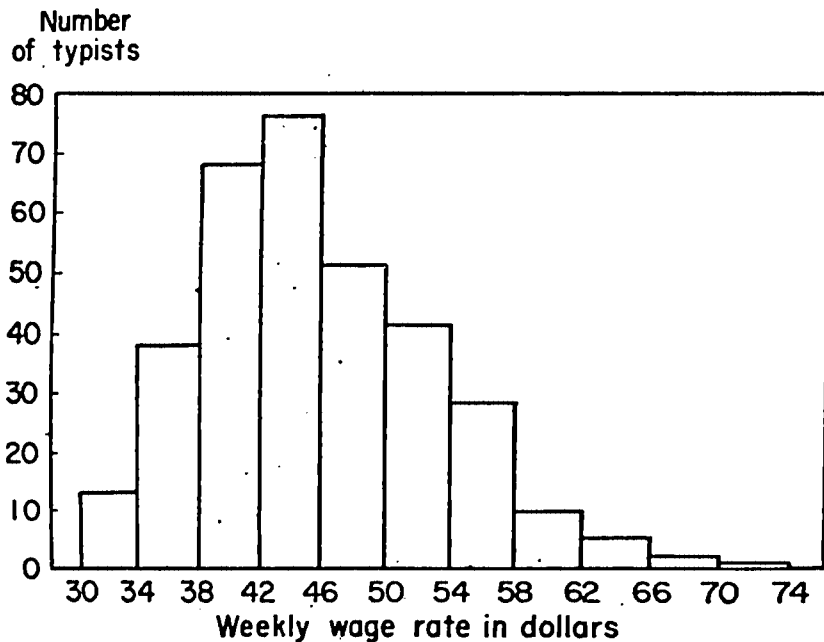


Chart 9.2. Histogram of Weekly Wage Rates of Senior Copy Typists in New York City, April 1949.

Source: *Studies in Labor Statistics*, No. 2, National Industrial Conference Board, Clerical Salary Survey of Rates Paid, April 1949, pp. 10-11.

figures for each class constitutes what is called the *histogram*. The total area of the histogram thus represents the total frequency as distributed throughout the classes.

How do we construct the histogram? All formal requirements—as to title, scale captions, and the like—are the same as for other arithmetic graphs. The *X*-axis, of course, need not start with zero, but the *Y*-axis must start with zero and must have no scale break. A space is left between the first rectangle and the vertical axis. The base of each rectangle is labeled on both sides in terms of the class limits if the data are continuous. In such a case, the upper limit of one class and the lower limit of the following class will coincide. In discontinuous data, only the lower limits are marked on the *X*-axis. Some statisticians, however, in representing discontinuous data leave small gaps between the rec-

tangles and label both limits of each class. Another way of labeling the horizontal axis on a histogram is by showing the mid value in the middle of the base of the rectangle.

An illustration of a histogram is found in Chart 9.2.

The histogram is distinguished from a bar chart in that it is two-dimensional; that is, in a histogram the width of the rectangles is a factor of importance. But what we visually compare in a histogram is often the height of the columns and not the area.

If the distribution has varying class intervals, we plot frequency densities. But an open-end class obviously cannot be plotted on a histogram; one solution is to plot the histogram without the open-end class or classes and to add the information concerning them in figures.

Frequency Polygon

"Polygon" literally means "many-angles." In statistics it means a curve representing a frequency distribution.

A frequency polygon may be looked upon as if it were derived from a histogram. If by straight lines we join the mid points of the upper horizontal sides of the rectangles in a histogram, we get a frequency polygon. But in actual construction we get the polygon by plotting for each class the value of its mid point against its frequency and joining by straight lines the points thus plotted.

A frequency polygon for the data shown in the histogram in Chart 9.2 is found in Chart 9.3.

Some statisticians favor closing the two ends of the polygon by continuing them to the base line. This procedure implicitly includes two hypothetical classes—one on each end of the distribution—each with a frequency of zero. The idea behind this extension is to make the area under the polygon equal to the area under the corresponding histogram.*

* "Smoothing" a polygon has special significance and is not to be done in every situation. It assumes that there is a basic "smoothed" form which the data would assume if we had a larger number of cases.

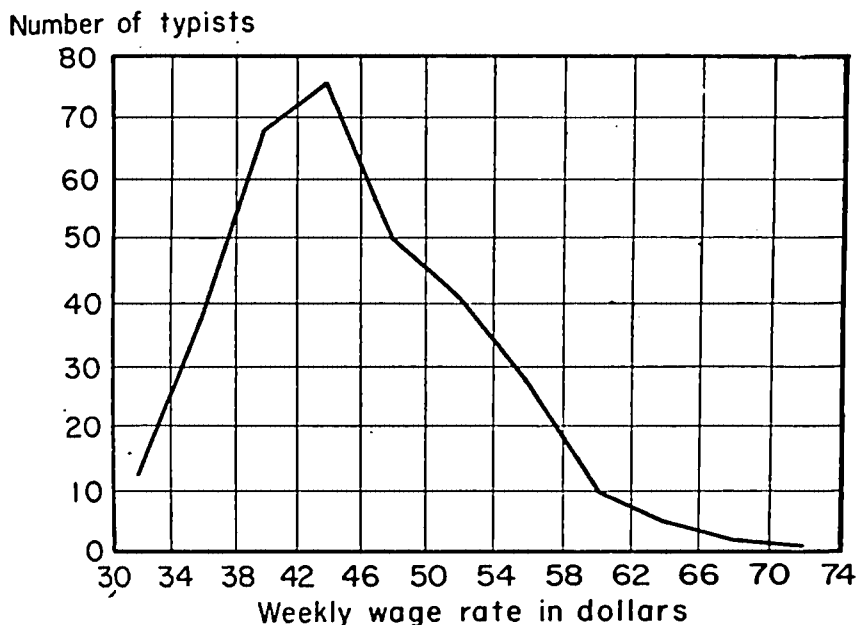


Chart 9.3. Frequency Polygon of Weekly Wage Rates of Senior Copy Typists in New York City, April 1949.

Source: Same as Chart 9.2.

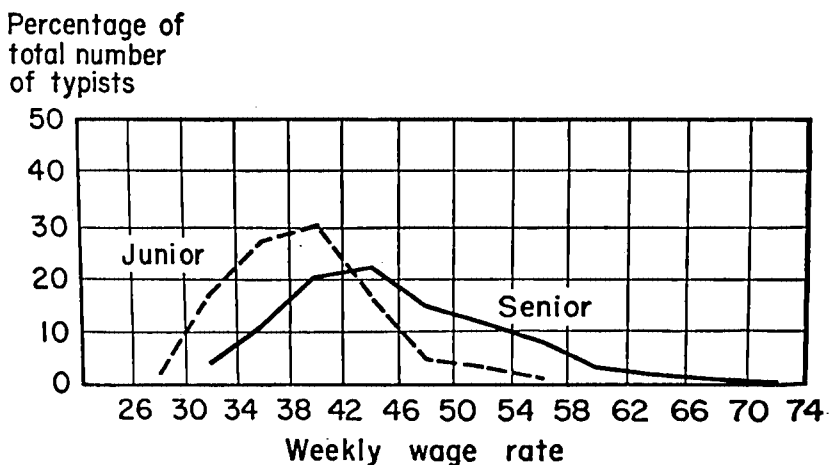


Chart 9.4. Percentage Distributions of Weekly Wage Rates of 120 Junior and 333 Senior Copy Typists in New York City, April 1949.

Source: Same as Table 8.5.

Formal requirements for presenting the polygon are the same as for the histogram. The problems of varying class intervals and open-end classes in the case of polygons are handled in the same way as in the case of histograms.

Two or more frequency polygons can be shown on the same graph; two histograms cannot. To compare histograms we must have a separate graph for each. Thus polygons are preferable for purposes of graphic comparison of frequency distributions.

To compare frequency distributions we usually have to use percentage frequencies. Accordingly, to compare frequency polygons we plot percentage frequencies. A comparison of polygons plotted in terms of percentage frequencies is shown in Chart 9.4.

Different Shapes of Frequency Polygons

Frequency distributions differ in their graphic shapes. From this variety of shapes certain basic types emerge. Some have their highest frequencies (which appear as the highest point on the graph) in the very center, with frequencies diminishing gradually as we go to the lower and the higher classes in value. Some have their highest frequencies at the very lowest values, in the lowest class; others have their highest frequencies in the class of highest value. Still others have their highest frequencies to the left, while others have them to the right of the class which is in the middle of the X values. Still others have two high frequencies, one at the lowest, the other at the highest.

In Illustration 9.1 are shown, schematically, the basic shape types of frequency polygons. Curve A is the type of what is called a bell-shaped curve, a symmetrical curve with the highest frequency in the central class and "tailing off" on each side in identical fashion. A special type of bell-shaped or symmetrical curve is the so-called "normal curve," whose significance will be seen later.

Curves B and C are types for *skewed* distributions. Skewness or asymmetry is the attribute of a frequency distribution that extends further on one side of the class with the highest fre-

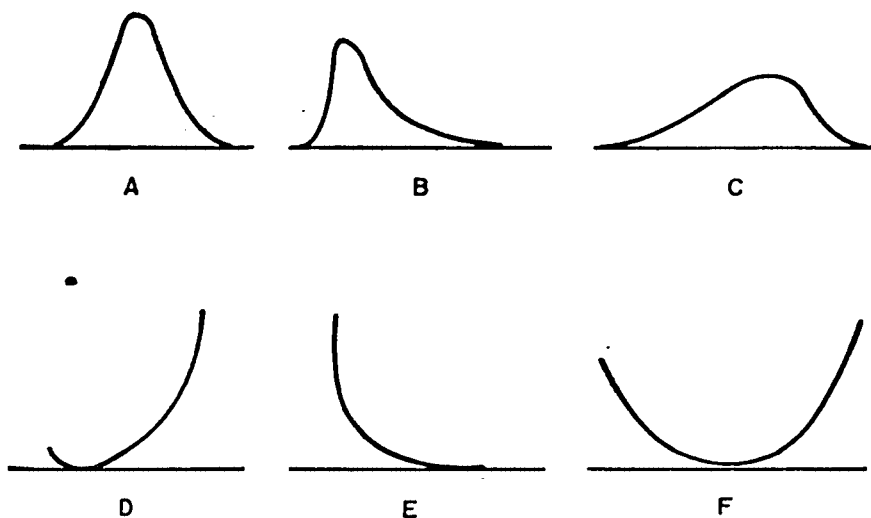


Illustration 9.1. Basic Shape Types of Frequency Polygons.

quency than on the other. A distribution can be either *skewed to the right* (positive skewness; curve B) or *skewed to the left* (negative skewness; curve C). A right-skewed distribution occurs when the curve has a tail to the right caused by high-value items in the distribution which are not compensated for by the presence of low-value items in the distribution; a left-skewed distribution, with a tail to the left, occurs when the curve is pulled towards low-value items which are not compensated for by high-value items in the distribution. These two skewed types are the most frequently found in economics and business, and the right-skewed more so than the left-skewed.

A frequency polygon which moves from low frequencies in low classes to its highest frequency in the highest class in the distribution, thus exhibiting its peak at the upper end of the distribution, is called a J curve, since it resembles the letter J (curve D). This is the type of curve which we get when we plot a frequency distribution relating to death rates by age but disregarding the death rate of young children.

A frequency polygon which has its highest frequencies at

the lowest values, gradually diminishing as we move toward the upper values, is called a reverse-J curve (curve E). Such a curve occurs when we plot a frequency distribution of bank accounts according to size, where the greatest number of depositors have the smallest accounts, and there is a gradual diminishing of frequencies as we move to the larger accounts.

If when we connect the plotted points on a graph of a frequency distribution we find that there are two high points, about equal in frequency, one at the lowest value and one at the highest value, we have what is known as a U curve, since it resembles the letter U (curve F). We arrive at this type of curve when we plot unemployment among employable males by age groups. Unemployment is highest among employable males at earliest working years and at latest working years. The U curve is also an example of a still larger class known as "bimodal curves," which have two peaks. (Bimodal distributions will be discussed later.) In the case of the U curve, the peaks are at the lower and upper extremes; in other cases of bimodal curves the peaks appear in other parts of the distribution.

The J curve, the reverse-J curve, and the U curve are quite unusual, and peculiar to given types of data.

Ogive

We have already discussed the cumulative frequency distribution in Chapter 8. Its graphic counterpart is the *cumulative frequency curve*, known as the *ogive*. The term ogive is taken from architecture where it refers to a diagonal rib of a vault, or to a pointed arch. An ogive portrays a distribution on a "less than" or an "or more" basis. From the ogive we can answer questions such as we mentioned in the treatment of the cumulative frequency distribution. In addition, the ogive can be used to locate certain measures graphically (see Chapter 11).

Along the X-axis of an ogive we plot one limit of each class. In a "less than" ogive—which is cumulated upward—we plot the upper limit of each class. In an "or more" ogive—which is

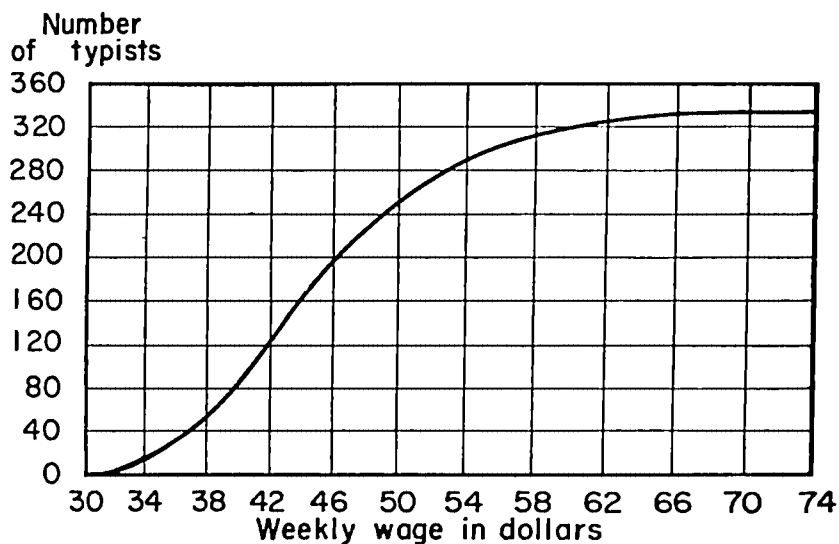


Chart 9.5. Senior Copy Typists Earning Less than Specified Weekly Wage, New York City, April 1949.

Source: Same as Chart 9.2.

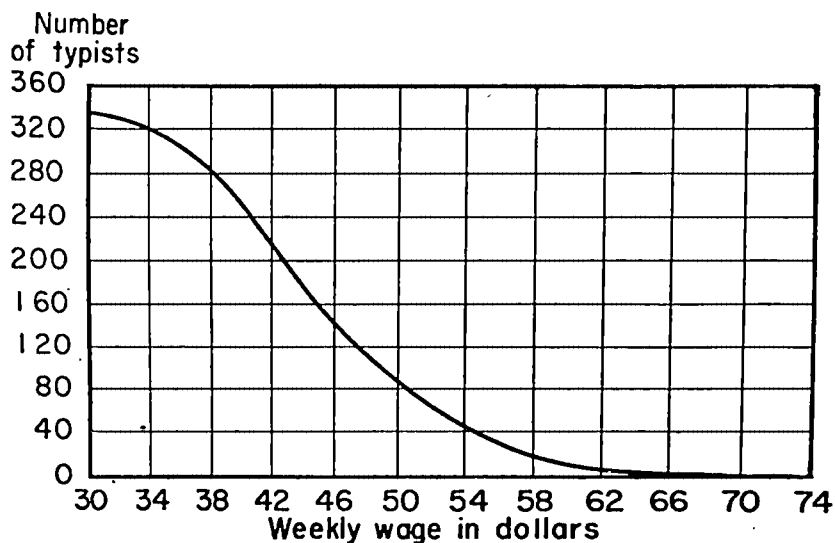


Chart 9.6. Senior Copy Typists Earning Specified Weekly Wage or More, New York City, April 1949.

Source: Same as Chart 9.2.

cumulative downward—we plot the lower limit of each class along the X -axis. Along the Y -axis we plot the *cumulative* frequency in each class.

A “less than” ogive is shown in Chart 9.5 and an “or more” ogive is shown in Chart 9.6.

Instead of cumulative frequencies in absolute numbers, cumulative percentage frequencies can be used on the Y -axis. These give a curve like the one plotted in terms of absolute frequencies. But we can then discover graphically what percentage of the cases in the distribution is less than a given magnitude, or what percentage of the cases is more than a given magnitude.

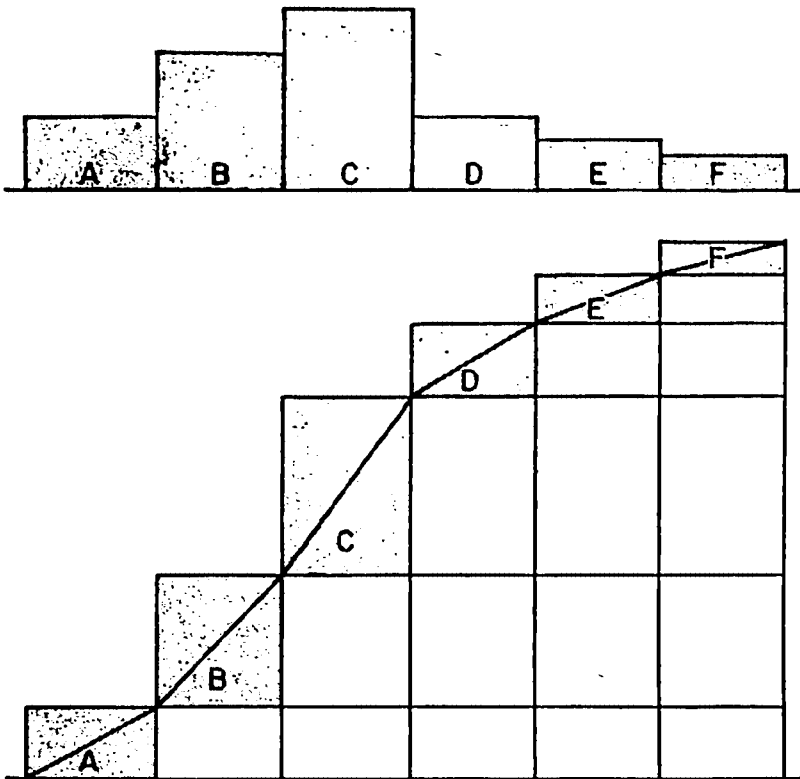


Illustration 9.2. Schematic Diagram Showing Derivation of Ogive from Histogram.

If the distribution has varying class intervals, the ogive may still be plotted with no difficulty. If the distribution is open-ended, no ogive should be plotted.

The data for the ogive of a given frequency distribution are the same data as for the frequency polygon (and histogram). But these data are arranged differently. Illustration 9.2 shows a schematic diagram of the derivation of the ogive (less than) from the histogram. The difference between the polygon and the ogive is the counterpart of the difference between the simple frequency distribution and the cumulative frequency distribution.

Summary

1. Arrays may be visualized in array charts. Frequency distributions may be presented graphically by means of histograms, polygons, and ogives.

2. Histograms present frequency distributions by rectangles of two dimensions, the widths signifying the class intervals and the heights the class frequencies.

3. The frequency polygon is a line diagram with class intervals on the X -axis and class frequencies on the Y -axis.

4. Histograms and polygons may be presented with frequencies expressed in absolute quantities or as percentages of the total frequency in the distribution.

5. Several type shapes may be distinguished in frequency polygons (the bell-shaped curve, skewed curve, J and reverse-J curves, U curve).

6. The ogive presents the cumulative frequency distribution graphically. The ogive may be on a "less than" basis or on an "or more" basis. Ogives may be presented in terms of cumulative absolute or percentage frequencies.

CHAPTER

10

Measurement of Masses: Averages: The Arithmetic Mean

Quantitative data in a mass exhibit certain general characteristics. (1) They show a tendency to concentrate at certain values, usually somewhere in the center of the distribution. Measures of this tendency are called measures of central tendency* or averages. (2) The data vary about a measure of central tendency. Measures of this deviation are called measures of variation or dispersion. (3) The data in a frequency distribution may fall into symmetrical or asymmetrical patterns. The measures of the direction and degree of asymmetry are called measures of skewness. (4) Polygons in frequency distributions exhibit peakedness. Measures of peakedness are called measures of kurtosis.

The purpose of these measures is to discover characteristics of mass data and hence to facilitate comparison within one mass or between masses of data.

Measures of central tendency or averages will be the subject of Chapters 10, 11, and 12; measures of variation, skewness, and kurtosis will be discussed in Chapter 13.

* This tendency toward centralization, though not universal, has established the expression "measure of central tendency" to describe an average. The term is imbedded in statistical language, but it is not always pertinent.

The Concept of Average

The concept of average is used constantly in everyday speech, and its everyday use gives some indication of its importance. "What kind of a worker is he?" it is asked. "Oh, about average," is the answer. "Do they pay high wages in that plant?" "Average wages."

What is the meaning of this concept? To say that a worker is an average worker means that he is typical of the group of which he is a part. To say that wages in a given plant are about average means that the wages paid in this plant are typical of wages paid in the industry.

In these cases, what is termed "average" is what is also called in statistics a *measure of central tendency*. A measure of central tendency is a typical value around which other figures congregate, or which divides their number in half. Thus, an average can be used to describe or represent a whole series of figures involving magnitudes of the same variable. That is, the average is an overall value. This measure permits us to compare individual items in the group with it and also permits us to compare different series of figures with regard to their central tendency.

There are several different kinds of averages; each one of them has certain characteristics, certain advantages, and certain disadvantages. We shall deal in this book with the four leading types of average, namely: (1) the arithmetic mean; (2) the median; (3) the mode; (4) the geometric mean.

THE ARITHMETIC MEAN

What the layman calls the "average" is in statistical terminology the arithmetic mean, which is only one of the types of statistical averages. The arithmetic mean is frequently referred to simply as the "mean"; and we talk about such values as mean income, mean tonnage, mean rental. As opposed to certain other averages which are found in terms of their posi-

tion in a series, the mean has to be computed by taking every value in the series into consideration. Hence, the mean cannot be found by either inspection or observation of the items. These and other characteristics will become clearer in later discussion of the mean and the other averages.

Mean from Ungrouped Data

The arithmetic mean is the quotient that results when the sum of all the items in the series is divided by the number of items.* For ungrouped data—that is, data not classified and arranged in a frequency distribution—the values are taken in their original form. Thus, if the items are 15, 18, 16, 14, 17, their sum is 80. The number of items is 5, and the mean is 16.

For general representation, each item in the series is given the symbol X ; the capital letter form is used.† Thus the income of one person, the weight of one aluminum casting, the age of one employee, in a series, is designated by X . The symbol for “the sum of” is the capital Greek letter sigma, which is Σ (called “capital sigma”) and which is read as “the sum of” whatever follows it. Thus ΣX means the sum of all the items in the X series. In our case, $\Sigma X = 80$. The symbol for the number of items in a series is N . In our case, $N = 5$. The symbol for the arithmetic mean is $\bar{X}$, that is, a capital X with a bar over it and is read “ X -bar.” In our case $\bar{X} = 16$.

Since the arithmetic mean for ungrouped data is the sum of all the items in the series divided by their number, the formula in terms of the above symbols is

$$\bar{X} = \frac{\Sigma X}{N},$$

* A per capita measure, such as the number of eggs consumed per person in the United States, is an arithmetic mean, since we divide in this instance (1) the total number of eggs consumed by (2) the total number of consumers, to get (3) the mean number of eggs consumed.

† Unfortunately, there is no universal agreement among statisticians as to the symbols to be used. The symbols here are the ones that seem to be most frequently used, but the student should be prepared to meet formulas using other symbols in statistical work.

where $\bar{X}$ = the arithmetic mean,

Σ = "the sum of,"

X = each individual item,

N = the number of items.

The arithmetic mean for ungrouped data can be worked out without an array; that is, from unarranged raw data.

The Weighted Mean

Up to this point we have had to give equal emphasis to each item in the series. This equal emphasis may be misleading if individual items have different importance, as in the following situation: The Smoke Shop sells Havana cigars at 25 cents, Manila cigars at 10 cents, and Wheeling cigars at 5 cents. What is the mean price? If the shop sells just 3 cigars, one of each, then $N = 3$, and $X = 25$, $X = 10$, $X = 5$, $\Sigma X = 40$, and

$$\bar{X} = \frac{\Sigma X}{N} = \frac{40}{3} = 13\frac{1}{3} \text{ cents.}$$

But the Shop actually sells 100 Wheelings, 60 Manilás, and 20 Havanas. Then our series of X 's is composed of three different-sized "bundles," the total of items in all three "bundles" being 180, and we can write

$$\begin{aligned}\bar{X} &= \frac{\Sigma(\text{Wheeling } X + \text{Manila } X + \text{Havana } X)}{(\text{Wheeling} + \text{Manila} + \text{Havana})} \\ &= \frac{\Sigma(100 \times 5¢ + 60 \times 10¢ + 20 \times 25¢)}{180}\end{aligned}$$

Now note those figures 100, 60, and 20. They are the quantities of the various classes of cigars sold; they are also the "weights," in statistical language, of the three prices. But note also that the sum of these weights is 180. Thus $180 = \Sigma w$, if we use w as a symbol for any weight, just as we used X as a symbol for any item. Similarly, we can write the sum of the items as $\Sigma(w \times 5¢ + w \times 10¢ + w \times 25¢)$, the three w 's being variously

valued, as we know. Moreover, since 5, 10, and 25 are all X , we can write the three "bundles" as one and we have ΣwX . Then

$$\bar{X} = \frac{\Sigma wX}{\Sigma w},$$

and in this equation we have the formula for the *weighted mean*. If we complete the computation, we proceed:

$$\bar{X} = \frac{\Sigma wX}{\Sigma w} = \frac{1600}{180} = 8.9 \text{ cents};$$

thus for the actual sales the mean price is a little under 9 cents. Table 10.1 summarizes the steps taken.

From this cigar example we can take two statements of principle: (1) weighting is designed to place the correct emphasis on each item according to its relative importance; (2) to apply a weight, we multiply an item in the series (here the cigar price) by the appropriate factor.

The weighted mean is particularly useful where the average we are looking for is a *mean of means*. If we have two arithmetic means, one from each of two series, involving the same variable of course, and we want to find the average for the two series, the arithmetic mean of each series cannot have the same

TABLE 10.1. WEIGHTED MEAN OF PRICES OF CIGARS SOLD BY THE SMOKE SHOP ON NOVEMBER 15, 1956.

Cigar	Price per cigar, cents X	Number sold, w	Price $\times$ weight wX
Wheeling	5	100	500
Manila	10	60	600
Havana	25	20	500
		180	1600

$$\Sigma w = 180, \quad \Sigma wX = 1600,$$

$$\bar{X} = \frac{\Sigma wX}{\Sigma w} = \frac{1600}{180} = 8.9.$$

weight as the other unless the number of items from which it was derived is equal to the number of items of the other. Hence, we weight each average by the number of items in its series, then add the two products thus obtained, and then divide the sum of these two products by the total number of items in both series together.

Thus, for instance, the arithmetic mean of the weekly wages in the Cosmos Manufacturing Corporation is \$45.00, while the arithmetic mean of the weekly wages in the Perfect Manufacturing Corporation is \$30.00. If we wanted to strike an average for both corporations, we would multiply each average by the number of workers in the corporation it represents. We then have the total payroll for each company. We add the two payrolls together, and divide by the total number of workers in both Cosmos Corporation and Perfect Corporation. This computation is worked out in Table 10.2.

TABLE 10.2. WEIGHTED MEAN OF MEAN WAGES OF COSMOS CORPORATION AND MEAN WAGES OF PERFECT CORPORATION, WEEK OF DECEMBER 8, 1945.

	<i>Cosmos Corporation</i>	<i>Perfect Corporation</i>
$\bar{X}$, wage in dollars.....	45	30
N , wage-earning units (workers).....	200	100
w	200	100
$w\bar{X}$	9000	3000

$$\Sigma w = 300, \quad \Sigma w\bar{X} = 12,000,$$

$$\bar{X} = \frac{\Sigma w\bar{X}}{\Sigma w} = \frac{12,000}{300} = 40.$$

We could perhaps have learned directly the total payroll data of the two companies and then have computed the combined mean without obtaining the individual means. We need the method, however, for such direct and pertinent data cannot always be had. Business firms and governments often report statistical items as ratios, means, or the like, concealing the original data. Then weights must be conjectured and used.

The computations are the same in principle. Suppose the mean wages of Cosmos and Perfect were known, \$45.00 and \$30.00 as before, but the companies refused to give the numbers of their employees. But a statistician might have good reason to believe that Cosmos has twice as many employees as Perfect, and assign to Cosmos a weight $w = 2$ and to Perfect a weight $w = 1$. Then

$$\bar{X} = \frac{\sum wX}{\sum w} = \frac{120}{3} = 40.$$

In averaging percents just as in averaging means we have to consider the absolute magnitudes to which they refer. For example, if we have information for the monthly production of men's and boys' sport shirts in some company, and the increase over the previous month's production is 20% in men's shirts and 50% in boys' shirts, then the mean increase in the production of both is not the simple mean of the percent increases, but their weighted mean. We weight each percent increase by the production in the previous month.

	<i>March Production, Dozens</i>	<i>April Production, Dozens</i>	<i>Percent Increase</i>
Men's shirts	800	960	20
Boys' shirts	<u>160</u>	<u>240</u>	50
All shirts	960	1200	

Thus the mean percent increase is

$$\frac{(20\% \times 800) + (50\% \times 160)}{960} = \frac{240}{960} = 25\%,$$

the denominator 960 being the March total.

In this way the weighting of each percent figure brings out the preponderant importance of the production of men's shirts.

We have had three examples of weighted means: (1) the mean prices of cigars sold by the Smoke Shop, (2) the combined mean wage of Cosmos and Perfect workers, and (3) the mean percent increase of shirt production.

In the first of these, we might have computed a mean from the prices alone, without reference to the numbers sold, thus:

$$\bar{X} = \frac{\Sigma X}{N} = \frac{40}{3} = 13\frac{1}{3}.$$

This mean would be the mean *price offered*, not the mean *price of cigars sold* (unless equal numbers of the three price classes of cigars were sold). The mean price offered would be of little probable use. Thus we illustrate the fact stated in Chapter 1, that a statistician must employ good judgment and good sense; mere proficiency in computation is not sufficient and may even develop false or misleading conclusions.

Averages and percentages cannot be treated as if they were original data. They are derived figures, and their importances are relative to the originals from which they are derived. The relative importance of each must therefore be brought out by the weighting. We illustrated this necessity in the examples of the mean of mean wages and the mean percent increase.

Mean from Grouped Data, Long Method

In a frequency distribution, we no longer have the original values of the items. Therefore, we have to deal with their representatives. Within each class, each item is assumed to have the value of the mid point of that class, as we have seen. The mid point has to be taken for each item in the class it represents. The mid point in each class is therefore multiplied by the class frequency. This gives us a series of products, one from each class. If these products are summed, we get a total similar to the total we would obtain from the original items; if the totals differ, the difference is due to the mid-point assumption. Just as in ungrouped data, we divide this total by the total number of items in the distribution to obtain the arithmetic mean.

The symbol for the arithmetic mean remains the same as for ungrouped data, namely $\bar{X}$. The symbol for the mid point of each class is capital X , the same as the symbol for the individual item in ungrouped data, on the assumption that each

item is now valued at the mid point. The symbol for "the sum of" always remains the same, namely Σ . But we need a new symbol, to represent "frequency"; this symbol is f . The total frequency in the distribution is symbolized by N , but can also be symbolized by Σf .

The formula for finding the arithmetic mean for grouped data is therefore as follows:

$$\bar{X} = \frac{\Sigma fX}{N}$$

This formula is for the arithmetic mean for grouped data by what is called the long method.

Let us work out the arithmetic mean for the frequency distribution of sales checks in the Alpha Store. This is found presented in Table 10.3.

TABLE 10.3. COMPUTATION OF ARITHMETIC MEAN BY LONG METHOD FOR FREQUENCY DISTRIBUTION OF SALES CHECKS IN THE ALPHA STORE IN DALLAS, TEXAS, ON SEPTEMBER 1, 1956.

Class	Mid point, X	Class frequency, f	Frequency multiplied by mid point, fX
\$1 and under \$3	\$2	3	\$6
\$3 and under \$5	\$4	9	\$36
\$5 and under \$7	\$6	25	\$150
\$7 and under \$9	\$8	35	\$280
\$9 and under \$11	\$10	17	\$170
\$11 and under \$13	\$12	10	\$120
\$13 and under \$15	\$14	1	\$14
		100	\$776

$$N = 100, \quad \Sigma fX = \$776,$$

$$\bar{X} = \frac{\Sigma fX}{N} = \frac{\$776}{100} = \$7.76.$$

The mean sale for the Alpha Store on September 1, 1956 is thus \$7.76. If we had found the arithmetic mean for the

original sales checks—that is, ungrouped data—the mean sale would be \$7.65. This difference of \$0.11 is due to the mid-point assumption, which as we have already acknowledged usually entails loss in accuracy.

Mean from Grouped Data, Short Method

There is a *short method* for computing the arithmetic mean: we guess a mean, and correct for the error in our guess. Although the short method does not save time and effort in ungrouped data, it is easier to grasp its essentials first in such data.

How do we correct for the error in our guess? The mean has an algebraic property on which this correction is based: the mean is a value such that the sum of the distances below it is offset by the sum of the distances above it. For example, the mean of 4, 5, 6, 7, 8 is 6. The sum of the differences of the items from the mean balance, that is, their algebraic sum, is zero, as follows:

$$\begin{array}{rcl}
 4 - 6 & = & -2 \\
 5 - 6 & = & \underline{-1} \quad -3 \\
 6 - 6 & = & \underline{+0} \\
 7 - 6 & = & +1 \\
 8 - 6 & = & \underline{+2} \quad \underline{+3} \\
 & & 0
 \end{array} \tag{1}$$

This property holds for the actual mean; it does not hold for the guessed mean unless the guess is correct. Thus, the algebraic sum of the differences of the items from the guessed mean will not be zero. For example, if in the same series 4, 5, 6, 7, 8 we guess 7 as the mean, we obtain the following:

$$\begin{array}{rcl}
 4 - 7 & = & -3 \\
 5 - 7 & = & -2 \\
 6 - 7 & = & \underline{-1} \quad -6 \\
 7 - 7 & = & \underline{0} \\
 8 - 7 & = & \underline{+1} \quad \underline{+1} \\
 & & -5
 \end{array} \tag{2}$$

What then is the correction necessary to adjust the guessed mean of 7? We take the average of these differences; that is, we divide -5 by N (which is 5) and get -1 . We add -1 and 7, the guessed mean, and this gives us 6, the actual mean. The minus sign in the sum of differences (-5) indicates that we guessed too high.

The difference of each item from the actual mean is symbolized by x , while the difference of each item from the guessed mean is symbolized by d . Hence $\Sigma x = 0$ as shown in example (1), and $\Sigma d = -5$ as shown in example (2).

On this foundation rests the short method for finding the mean for grouped data. Theoretically, in grouped data, too, any value may be guessed as the mean, but in practice it is useful to guess one of the mid values. It does not matter which mid value is taken as the guessed mean.

For the frequency distribution in Table 10.3, let us guess the mean as the mid value of the class "\$7 and under \$9," or \$8. The second step consists of correcting the guessed mean.

Instead of taking the differences between the guessed mean and each individual item, as in ungrouped data, we take the differences between the guessed mean and the representatives of the individual items, or the mid points of each class.

But a further saving is possible. Instead of the actual differences between mid values and guessed mean, we can count the number of classes that separate each mid point from the guessed mean. Thus, obviously the mid value of the "\$7 and under \$9" class is not separated from the guessed-mean class and its difference is therefore 0.

The mid value of the "\$5 and under \$7" class (or \$6) is one step *lower* than the guessed mean and therefore its difference in terms of steps is -1 . The mid value of the "\$13 and under \$15" class is three steps *above* the guessed mean and therefore its difference in terms of steps is $+3$. These are called step deviations and are shown in Table 10.4. They are designated by the symbol d .

The step deviation for each class is taken as many times as

TABLE 10.4. COMPUTATION OF ARITHMETIC MEAN BY SHORT METHOD FOR FREQUENCY DISTRIBUTION OF SALES CHECKS IN THE ALPHA STORE IN DALLAS, TEXAS, ON SEPTEMBER 1, 1956.

Class	Mid point of class, X	Frequencies in class, f	Step deviation from class of guessed mean, d	Step deviation times frequency, fd
\$1 and under \$3	\$2	3	-3	-9
\$3 and under \$5	\$4	9	-2	-18
\$5 and under \$7	\$6	25	-1	-25
\$7 and under \$9	\$8 = $\bar{X}_d$	35	0	0
\$9 and under \$11	\$10	17	+1	+17
\$11 and under \$13	\$12	10	+2	+20
\$13 and under \$15	\$14	1	+3	+3
Total		100		-12

$$N = 100, \quad \Sigma fd = -12;$$

$$\begin{aligned}
 \bar{X} &= \bar{X}_d + \left(\frac{\Sigma fd}{N} \right) i \\
 &= \$8 + \left(\frac{-12}{100} \right) \$2 \\
 &= \$8 - \$0.24 \\
 &= \$7.76.
 \end{aligned}$$

there are items in the class. Therefore we multiply d for each class by its class frequency or f . This procedure gives a column of products symbolized by fd . Again we average the deviations. We sum the column of fd (in our case $\Sigma fd = -12$) and divide by the total number of items in the distribution. In our case $N = 100$. This process gives a correction of -0.12 .

But $\frac{\Sigma fd}{N}$, or in our case -0.12 , is not in dollars, but in step deviations. We have up to now neglected the width of the class. We must reintroduce the size of the class interval, symbolized

by i , and transform the correction of -0.12 into dollars. Hence we multiply -0.12 by the class interval $i = \$2$. This gives us a correction factor of $-\$0.24$.

We subtract $\$0.24$ from the guessed mean or $\$8$ and obtain the actual mean of $\$7.76$. The guessed mean is symbolized by $\bar{X}_d$.

We have thus employed the following formula for finding the arithmetic mean for grouped data by the short method:

$$\bar{X} = \bar{X}_d + \left(\frac{\sum fd}{N} \right) i.$$

The mean found by the short method for grouped data is identical with the mean found by the long method for grouped data. But it too differs from the mean found for corresponding ungrouped data.

If we guess a mean too low, the correction factor comes out positive, as is shown in Table 10.5. But we arrive at the same actual mean.

It should now be clear that the figures used in the short method are much simpler and that computation therefore is much easier. This will be found most strikingly true in long distributions with large class intervals.

Where the class intervals are not uniform, the short method of computing the mean cannot be used, but the long method for grouped data may be used.

Characteristics of the Mean

We are now in a position to understand certain of the characteristics inherent in the arithmetic mean.

In the computation of the mean every item counts. As a result, extreme values at either end of the series severely affect the value of the mean. For example, we have the net sales of a large number of corporate enterprises in the United States. We want to find their average net sales. The inclusion of gigantic corporations like General Motors and General Electric and United States Steel will distort the picture. The arithmetic

TABLE 10.5. COMPUTATION OF ARITHMETIC MEAN BY SHORT METHOD (GUESSING MEAN IN LOWEST CLASS) FOR FREQUENCY DISTRIBUTION OF SALES CHECKS IN THE ALPHA STORE, DALLAS, TEXAS, ON SEPTEMBER 1, 1956.

Class	Mid point of class, X	Frequencies in class, f	Step deviation from class of guessed mean, d	Step deviation times frequency, fd
\$1 and under \$3	\$2 = $\bar{X}_d$	3	0	0
\$3 and under \$5	\$4	9	+1	+9
\$5 and under \$7	\$6	25	+2	+50
\$7 and under \$9	\$8	35	+3	+105
\$9 and under \$11	\$10	17	+4	+68
\$11 and under \$13	\$12	10	+5	+50
\$13 and under \$15	\$14	1	+6	+6
Total		100		+288

$$N = 100, \quad \Sigma fd = +288;$$

$$\begin{aligned}
 \bar{X} &= \bar{X}_d + \left(\frac{\Sigma fd}{N} \right) d \\
 &= \$2 + \left(\frac{288}{100} \right) \$2 \\
 &= \$2 + \frac{\$576}{100} \\
 &= \$7.76.
 \end{aligned}$$

mean would be pulled upwards as a result of the extreme values of the net sales of these three giants. This average would make it appear that the typical net sales of all the corporations in the series was very high.

The mean may be characterized as a point of balance. The sum of the positive deviations from it is equal to the sum of the negative deviations from it. This equivalence gives us the *first* mathematical characteristic of the mean: that is, the sum

of all deviations, taking signs into account, is zero. Or, as we have seen on page 137, in symbols this is expressed as $\Sigma x = 0$.

A *second* mathematical characteristic of the mean is that the sum of the squared deviations of the items from the mean is a minimum, that is, less than the sum of the squared deviations of the items from any other value. For example, if we have a series of values 4, 5, 6, 7, 8, the mean is 6. The squared deviations are as follows:

$$\begin{array}{rcl} 4 - 6 = -2, & -2^2 = 4 \\ 5 - 6 = -1, & -1^2 = 1 \\ 6 - 6 = 0, & 0^2 = 0 \\ 7 - 6 = +1, & +1^2 = 1 \\ 8 - 6 = +2, & +2^2 = 4 \\ & \Sigma x^2 = 10 \end{array}$$

If we square the deviations from some other value, their sum must be larger than 10. For example, let us find the sum of the squared deviations of the items here from 4 (symbolizing the deviations in this computation by d):

$$\begin{array}{rcl} 4 - 4 = 0, & 0^2 = 0 \\ 5 - 4 = +1, & +1^2 = 1 \\ 6 - 4 = +2, & +2^2 = 4 \\ 7 - 4 = +3, & +3^2 = 9 \\ 8 - 4 = +4, & +4^2 = 16 \\ & \Sigma d^2 = 30 \end{array}$$

Hence, Σx^2 is a minimum. The importance of this characteristic will be seen later in our discussions of the standard deviation, and of the least-squares method in time series and correlation.

A *third* mathematical characteristic of the mean is that if we replace each item in a series by the mean, then the sum of these substitutions will equal the sum of the individual items. Thus, in the figures 4, 5, 6, 7, 8, if we substitute the mean for each item—here 6, 6, 6, 6, 6—the total is 30 both in the original

series and in the substitution series. Expressed in symbols, this gives the following:

$$\text{Since } \bar{X} = \frac{\Sigma X}{N}, \text{ therefore } N\bar{X} = \Sigma X.$$

Hence, if we have the mean number of workers for industrial plants in a city, $\bar{X}$, and if we know the number of industrial plants in this city, N , then we can arrive at the total number of industrial workers in the city, ΣX .

Summary

1. "Central tendency" is one of the four aspects of frequency distributions that can be measured.
2. The arithmetic mean is one of the measures of central tendency or averages.
3. The arithmetic mean is the sum of all the items divided by their number.
4. The arithmetic mean can be found from raw data, from data in an array, and from grouped data.
5. In a weighted mean, we take account of the relative importance of items. In averaging means and in averaging percentages, we have to use a weighted mean.
6. There is a long method and a short method of finding the mean. Both methods give the same answer.
7. In computing the arithmetic mean the value of every item counts. Thus, extreme values influence the mean strongly.
8. The mean has the following mathematical properties: $\Sigma x = 0$; $\Sigma x^2 = \text{a minimum}$; $N\bar{X} = \Sigma X$.

CHAPTER

11

Measurement of Masses: Averages: The Median; the Mode; the Geo- metric Mean

THE MEDIAN

The Concept of the Median

The Federal Reserve Board in a study of 1948 family incomes found that the mean money income per family was approximately \$3600 a year, but about as many families received less than \$3000 in cash income as received more than \$3000. As distinct from the mean, which here is \$3600, we are faced with a different type of average, which here is \$3000. This average is a value which in this case separates incomes in the United States into two equally large groups. This type of average is called the median.

As distinct from the arithmetic mean, which is calculated from the *value* of every item in the series, the median is what is called a *position* average. The term "position" refers to the place of a value in a series. The position of the median in a series is such that the number of items (in the series) below it

TABLE 11.1. FINDING THE MEDIAN FOR UNGROUPED DATA FROM ARRAY OF PRODUCTIVITY RATES OF INDIVIDUAL WORKERS IN THE COSMOS CORPORATION, JUNE 30, 1956.

<i>Item Number</i>	<i>Productivity Rate</i>
1	24.50
2	25.25
3	25.50
4	25.50
5	27.00
6	28.75
7	29.00
8	→ 29.00
9	30.25
10	30.75
11	31.25
12	32.50
13	34.00
14	35.00
15	36.25

The series has 15 items ($N = 15$). The middle item is the eighth item. The eighth item is 29.00, therefore the median is 29.00.

in magnitude equals the number of items (in the series) above it in magnitude. Thus, the median is a value in a series which is exceeded by as many values as it exceeds.*

The Median for Ungrouped Data

In order to find the median position in ungrouped data, an *array* must be made. The middle value in any haphazard arrangement has no meaning as a measure of central tendency since it may have larger as well as smaller items on both sides of it.

In the series of figures 2, 3, 4, 5, 8, the median is 4. In this series 4 is the magnitude such that the number of items lower than 4 is equal to the number of items higher than 4. If the series has an even number of items, such as 1, 2, 3, 4, 5, 8, no

* A median may be surrounded by neighboring values that are equal to it. Thus in the series 3, 5, 6, 7, 7, 7, 9, 11, 12 the median is 7.

one of these figures by itself divides the series in half. In an even-numbered series, we assume the median to be halfway between the two middle items—here 3 and 4—and the median therefore is $3\frac{1}{2}$.

To determine the median for ungrouped data if the series has an odd number of items, we do the following:

1. Make an array of the raw data.
2. Count the items and find the middle item.
3. Take the value of this middle item as the median.

The median for an odd-numbered ungrouped series is worked out in Table 11.1.

If the array has an even number of items, there is no actual value exactly in the middle of the series. Thus, if there are 24 items in a series, the median *position* is 12.5; that is, the median value is halfway between the value of the items that are 12th and 13th in order of magnitude.

The median for an even-numbered ungrouped series is worked out in Table 11.2.

TABLE 11.2. FINDING THE MEDIAN FOR UNGROUPED DATA FROM ARRAY OF PERCENT SCORES IN GENERAL APTITUDE TEST FOR INDIVIDUAL WORKERS IN THE PERFECT CORPORATION, JANUARY 15, 1957.

Item number	Percent score	Item number	Percent score	Item number	Percent score
1	32	9	58	17	78
2	41	10	62	18	79
3	45	11	70	19	82
4	49	12	70	20	83
5	50	13	→ 73 ←	21	84
6	50	14	74	22	87
7	51	15	75	23	90
8	54	16	76	24	98

The series has 24 items ($N = 24$). The median is between the twelfth and thirteenth items.

The twelfth item is 70, the thirteenth item is 73, therefore the median is 71.5.

The Median for Grouped Data

In grouped data, the individual items have lost their identity, and the middle item cannot be found by counting. It is necessary to get inside of a class to find the value that divides the number of all items in half. If we divide the number of frequencies (N) in two halves, we find that the middle item falls within a class. Which class? To establish this class we cumulate frequencies until we reach the lowest class whose cumulative frequency is greater than $\frac{N}{2}$, commonly written $N/2$. This class is called the median class.

But at what value in the median class does the median fall? We have not reached $N/2$ in our cumulative frequencies when we enter the median class. Assuming that all items are evenly distributed over this median class, we proceed toward the upper limit of this class, stopping when we have picked up our missing frequencies. This operation brings us to a value within the median class which is presumed to have $N/2$ items on each side of it.

How do we find this value? First of all, it is at least as high as the lower limit of the median class and may be higher. If higher (as is usually the case), how much higher? Higher by a proportion of the number of items we are short when we enter the median class, to the frequency (total number of items) in the median class. To find this proportion, we divide the number of items we are short to make up $N/2$ by the number of all the items in the median class. To transform this fraction into an actual value in the series, we multiply it by the size of the class interval of the median class. We add the result of this interpolation to the lower limit of the median class. This sum gives us the median.

The steps for finding the median for grouped data are therefore as follows:

1. Divide the number of items in the distribution by 2; that is, compute the value of $N/2$.
2. Accumulate frequencies.

3. Find the class whose cumulative frequency is the first to exceed $N/2$.^{*} This is the median class.
4. Find the *actual* lower limit of the median class.
5. Then perform the following operations: Subtract from $N/2$ the frequencies we have accumulated before entering the median class. Divide this difference by the frequency in the median class. Multiply the quotient thus obtained by the size of the class interval of the median class.
6. Add the result of the operation in step 5 to the lower limit of the median class. This sum gives us the median.

The formula for this procedure is

$$\text{median} = l_1 + \left(\frac{\frac{N}{2} - \Sigma f_1}{f_{\text{med}}} \right) i,$$

where l_1 = the lower limit of the median class,^{*}

N = the total frequency,

Σf_1 = the sum of all frequencies accumulated before entering the median class,[†]

f_{med} = the frequencies in the median class, and

i = the size of the class interval of the median class.

The median for a frequency distribution is found worked out in Table 11.3. A schematic diagram showing the median value is given in Illustration 11.1.

The median of 127.8 workers found in Table 11.3 means that one-half of the industrial establishments in this city have less than this number of workers and one-half have more. In cases such as this, the median may turn out to be a value which cannot appear in the series. The median 127.8 workers is therefore an abstraction.

^{*} It may happen that the cumulative frequency of a class equals $N/2$. In such a case, this class is the median class and its upper limit is the median of the distribution.

[†] The subscript 1 in this book designates "the preceding" and subscript 2 designates "the following." Thus, l_1 means the limit of the median class bordering the preceding class.

The median can also be found for grouped data by entering the median class at its upper limit. In this case the formula is

$$\text{median} = l_2 - \left(\frac{\frac{N}{2} - \Sigma f_2}{f_{\text{med}}} \right) i,$$

where l_2 = the upper limit of the median class,*

TABLE 11.3. FINDING THE MEDIAN FOR SIZE OF INDUSTRIAL ESTABLISHMENTS BY NUMBER OF WORKERS IN THE CITY OF OMEGA, JULY 1, 1956.

Number of workers	Number of establishments (= frequencies)	Cumulative frequencies
0 to 49	46	46
50 to 99	59	105
100 to 149	↔ 45	150
150 to 199	37	
200 to 299	31	
300 to 399	20	
400 to 499	13	
500 and over	9	
	—	
	260	

Arrow indicates median class.

$$N = 260,$$

$$\frac{N}{2} = 130,$$

$$l_1 = 100$$

$$\Sigma f_1 = 105,$$

$$f_{\text{med}} = 45,$$

$$i = 50.$$

$$\text{median} = l_1 + \left(\frac{\frac{N}{2} - \Sigma f_1}{f_{\text{med}}} \right) i$$

$$= 100 + \left(\frac{130 - 105}{45} \right) 50$$

$$= 100 + \left(\frac{25}{45} \right) 50$$

$$= 127.78 \text{ workers}$$

or 127.8 workers.

* In discrete data, use the lower limit of the next higher class.

N = the total frequency,

Σf_2 = the sum of all the frequencies accumulated from the highest class to the class immediately above the median class in value,

f_{med} = the frequencies in the median class,

i = the size of the class interval of the median class.

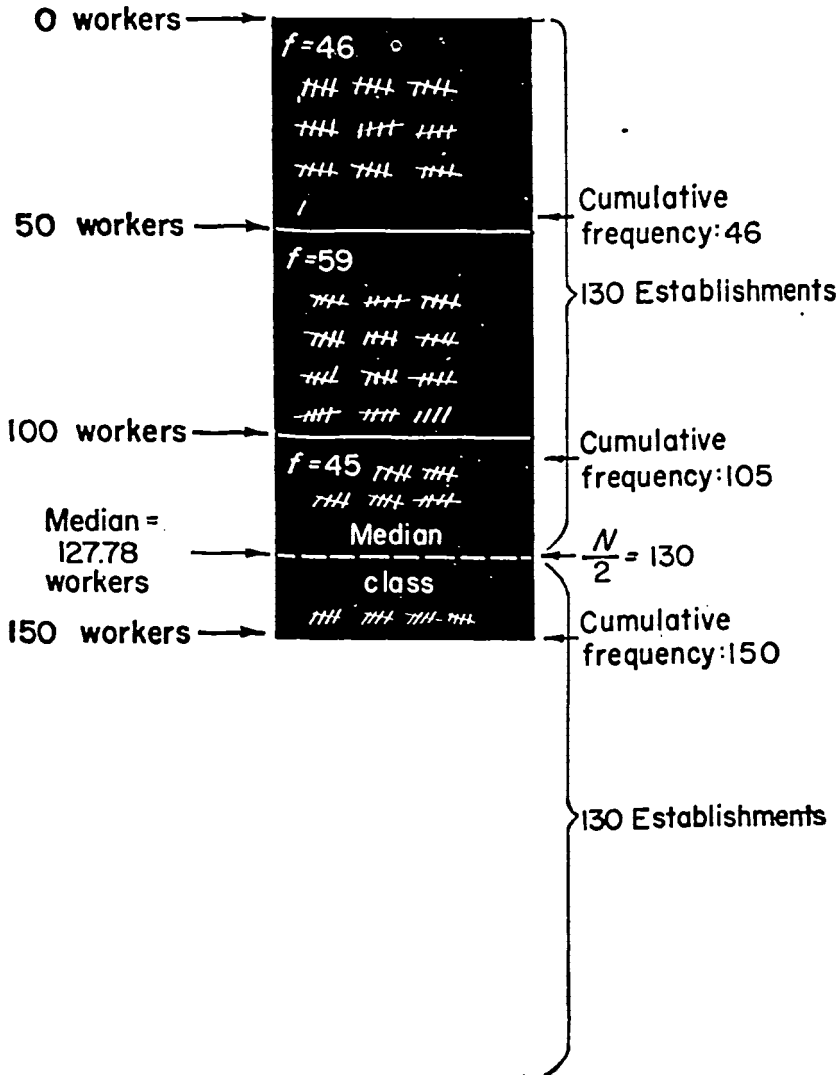


Illustration 11.1. Schematic Diagram for Median of the Frequency Distribution in Table 11.3.

Characteristics and Uses of the Median

The median is in a sense also a point of balance. The mean balances *differences from itself*; the median balances *numbers of items*; that is, a series arranged in order of magnitude is divided by the median into two equal parts. In a series graphically represented, a cut through the frequency polygon at the median value separates it into two equal areas.

As will be seen later, there is a mathematical property of the median which is important in finding certain measures of variation: the sum of the differences of each of the items in a series from the median is a minimum, if signs are ignored. Thus, in the series of 4, 5, 9, 11, 14, the median is 9. The differences, disregarding sign, are 5, 4, 0, 2, 5. Their sum is 16. This sum is smaller than the sum of the differences between each of the items in the series and any value other than the median, if we disregard signs. For instance, the deviations from 5 are 1, 0, 4, 6, 9, and their sum is 20. This mathematical property of the median is symbolized thus: $\Sigma|x| = \text{a minimum}$; the bars around $|x|$ mean "disregarding signs."

In Table 11.3 we have a distribution that is open-ended and has varying class intervals. In such a situation the median is especially useful. Moreover, in markedly skewed distributions, such as income distributions, the median is very often used.

Let us take an additional example. Suppose we want to set up two production lines with an equal number of workers on each. We do not want fast workers and slow workers on the same line, since the fast workers would swamp the slow workers and the slow workers would retard the fast workers. But each worker has a productivity rate, which we have learned through time-and-motion studies. If we selected a productivity rate that divides the group in half, so that one-half is below this productivity rate and one-half above, we can set up two production lines, one for slow workers and one for fast. This efficient assignment is made possible by finding the median value.

A special feature of the median is illustrated here, namely

that it is the most appropriate average in dealing with rates, ranks, and other types of items that are not counted or measured, but are scored.

Related Positional Measures

There are other measures which divide a series into equal parts. Of chief importance are the quartiles, the deciles, and the percentiles. Three quartiles divide a series into four equal parts, nine deciles into ten equal parts, ninety-nine percentiles into one hundred equal parts. The median, it should be clear, is the same value in a series as the second quartile, the fifth decile, and the fiftieth percentile.

In economics and business, the quartiles are more widely applied than deciles or percentiles. The first quartile, Q_1 , is a value such that 25% of the items in the series are equal to or smaller than it, while the third quartile, Q_3 , is a value such that 75% of the items in the series are equal to or below it. Since the quartiles are used to find one of the measures of variation, we shall discuss their computation in that place (Chapter 13, p. 176 ff.).

The deciles and percentiles are important in psychological and educational statistics concerning grades, rates, scores, and ranks; they have bearing in economics and business statistics in personnel work, productivity ratings, and other such situations.

THE MODE

The Concept of the Mode

When people talk about the "average consumer," for example, they usually mean the type of consumer who is met *most frequently* with regard to expenditures or some other quality. The consumer expenditure *most frequently* met with is known in statistics as the *modal* expenditure. Thus, the mode is the most common value, the value which occurs most fre-

quently in a series. It is the most easily understood of the main types of average; thus the modal retail price paid for an electric toaster, for example, is the retail price paid for the commodity more often than any other price. Hence, if you are going to buy an electric toaster, the statistical chances are highest that you will buy one at the modal price. Therefore, the modal value may be looked upon as the value in a series most likely to occur.

The Mode for Ungrouped Data

To find the mode for ungrouped data is a simple matter. All we have to do is find the value that occurs most frequently. This statement assumes, of course, that there is a noteworthy repetition of values somewhere in the series. To discover such repetition we must first make an array. For instance, in the series 2, 4, 5, 5, 5, 8, 9, 10 the mode is 5. But even in an array, we usually do not find noteworthy repetition that can be called a typical value. If we have the population of every city in the United States arranged in order of magnitude, we do not find repetition of values. But if we group the same data, a modal population for United States cities will appear.

The Mode for Grouped Data

It is very simple to find the modal class by inspection. The modal class is the class with the highest frequency. In a histogram, the modal class is the class with the highest column. But then we do not have a single representative value for the series, but a range of values. However, the value range of the modal class may on occasion suffice for practical purposes.

In most cases, we wish to find the modal value within the modal class. But to find the "true" mode advanced methods are required. With the tools available to us in basic statistics, only a "crude" mode can be obtained. Thus, in basic statistics, we can arrive only at an estimate of the mode.

The mid point of the modal class may be used as a rough estimate of the mode. This practice assumes that the modal

class has a class on each side of it of equal strength and pull; that is, that the mode is not being pulled up or down from the mid point of the modal class. But actually, in most frequency distributions the modal class is flanked by neighbors of unequal strength; that is, the premodal class may have fewer or greater frequencies than the postmodal class. This imbalance requires us to assume that the frequencies in the modal class are distributed unevenly (see Illustration 11.2). Hence, we correct

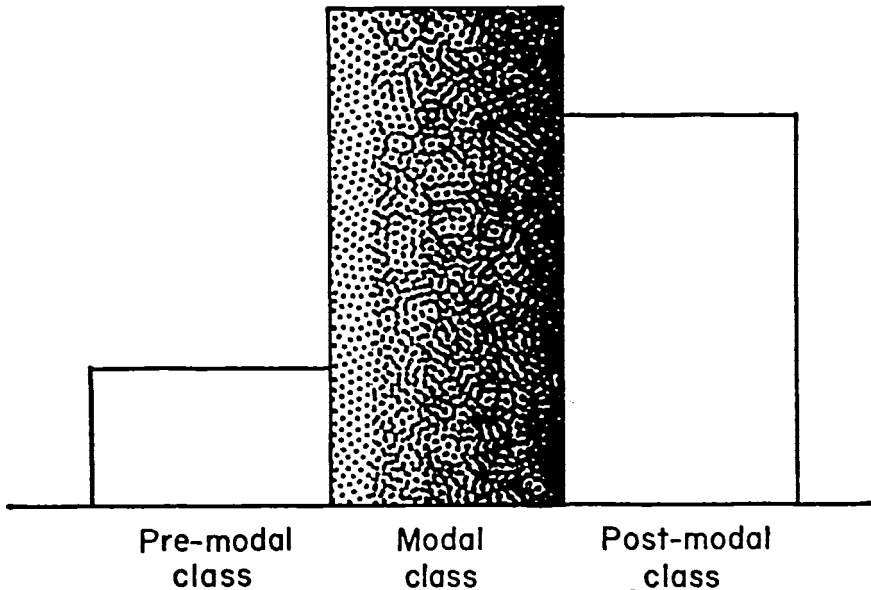


Illustration 11.2. Uneven Distribution of Items within a Modal Class.

the mid point of the modal class in the direction of the neighboring class with the higher frequency. This is the basis for finding the “crude” mode for grouped data.*

We must go into the modal class at its lower limit. The detailed steps illustrated in Table 11.4 are as follows:

1. Find the class with most frequencies; this is the modal class.
2. Establish the *actual* lower limit of this class.

* Another method for estimating the mode will be discussed in the next chapter when we compare the three averages.

TABLE 11.4. FINDING THE MODAL WEEKLY INCOME OF PART-TIME WORKERS IN THE N. & M. STORES, DES MOINES, IOWA, JUNE 30, 1956.

<i>Weekly Income, Dollars</i>	<i>Number of Workers</i>
20 and under 30	85
30 and under 40	120
40 and under 50	110
50 and under 60	67
60 and under 70	49
70 and under 80	21
80 and under 90	6
	458

$$l_1 = 30, \quad \Delta_1 = 35, \quad \Delta_2 = 10, \quad i = 10;$$

$$\begin{aligned} \text{mode} &= l_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) i = 30 + \left(\frac{35}{35 + 10} \right) 10 \\ &= 30 + \left(\frac{35}{45} \right) 10 = 30 + 7.78 \\ &= \$37.78. \end{aligned}$$

3. Subtract the number of frequencies in the premodal class from the number of frequencies in the modal class. The result is symbolized by the Greek letter, capital Δ (delta) with a subscript 1, thus: Δ_1 (Δ stands for "difference").

4. Subtract the number of frequencies in the postmodal class from the number of frequencies in the modal class. This result is symbolized by Δ with a subscript 2, thus: Δ_2 .

5. Divide Δ_1 by the sum of Δ_1 and Δ_2 , and multiply the quotient obtained by the size of the class interval of the modal class.

6. Add the result obtained by performing the operation in step 5 above to the lower limit of the modal class. This sum is the mode.

The formula for this procedure (difference method) is:

$$\text{mode} = l_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} i,$$

where l_1 = the lower limit of the modal class,

Δ_1 = the difference between the frequencies in the modal class and the premodal class,

Δ_2 = the difference between the frequencies in the modal class and the postmodal class, and

i = the size of the class interval of the modal class.

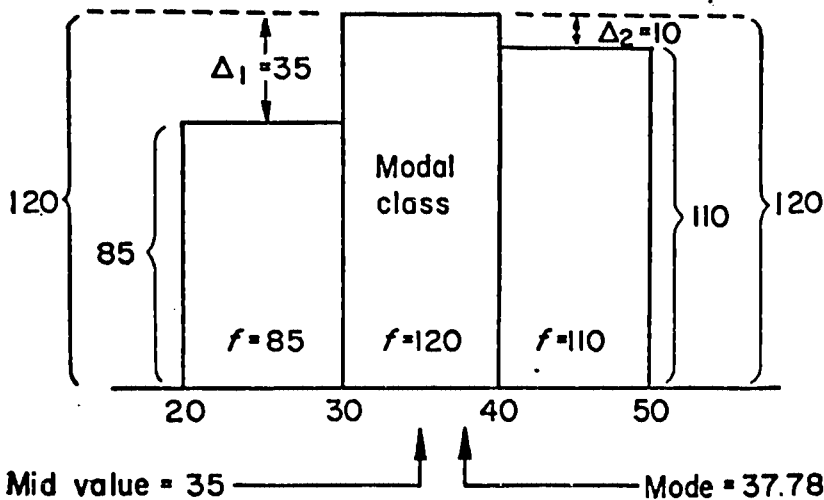


Illustration 11.3. Schematic Diagram for the Crude Mode of the Frequency Distribution in Table 11.4.

Illustration 11.3 is a schematic diagram visualizing the basis for finding the crude mode.

The mode of \$37.78 found in Table 11.4 means that in the N. & M. Stores more part-time workers receive *approximately* \$37.78 than any other wage.*

Special Problems of the Mode

The concept of the highest frequency, the mode, is dependent upon the classification system, and varying the size of the interval within the distribution (varying the class intervals) or throughout the distribution (making a uniform interval larger or smaller) usually results in a shifting of concentration of frequencies, and therefore a shifting of the crude mode. Moreover, the mode

* We stress here the approximate nature of this crude mode because the value found by this formula may actually not be the most frequent value and may not even appear in the series.

is not informative in distributions which have their concentrations in the lowest class or in the highest class, or where two or more separate and distinct concentrations occur. In general, in such situations there is no meaningful measure of central tendency at all.

Where two separate and distinct concentrations of frequencies occur in a frequency distribution, we have what is called *bimodality*. What causes bimodality?

1. The data may be heterogeneous. Let us return to the illustration on bank accounts given in Chapter 8 on page 95. If we do not separate individual bank accounts from corporate bank accounts, we get a concentration around a low value for individual accounts and around a high value for corporate accounts.

2. The data may be poorly grouped. Too small a class interval may produce bimodality.

3. Mere chance may produce bimodality.

4. The data may be inherently bimodal even though homogeneous.

We talk of bimodality even though the two concentrations are not equal, but are distinct. If the same highest frequency, however, occurs in two adjacent classes, the condition is not bimodality. In such a case, we consider the modal class to consist of both classes together, and we base our computation of the mode on the class interval of the combined modal class.

GRAPHIC ANALYSIS

By the use of graphs we can find the median and the mode. To be sure, the precision of the result thus found is no greater than the precision of the graphic technique and the clarity with which the graph can be read,

Median and Related Measures

The measures discussed in this chapter up to this point can be *obtained* through graphs. Graphs are thus used here for

analysis and not for presentation only. These measures are the median and related positional measures, and the mode.

In computing the median we worked through cumulative frequencies. The graph of a cumulative frequency distribution is, as we have seen, an ogive. The median is the value corresponding to the point where the cumulative frequency is half of the sum of the total frequency, or $N/2$. Thus, to find the median on an ogive, we first locate $N/2$ on the Y -axis which shows the cumulative frequencies. We then find the corresponding point in the ogive by drawing a horizontal line from $N/2$ on the Y -axis to the ogive. At the point where this horizontal line meets the ogive, we drop a perpendicular to the X -axis. Where this perpendicular meets the X -axis, we can read the median. This procedure is illustrated in Chart 11.1.

The same procedure followed in Chart 11.1 for graphically finding the median may be employed on an "or more" ogive.

The point from which the perpendicular is to be dropped to the X -axis may be found also by plotting both ogives on one

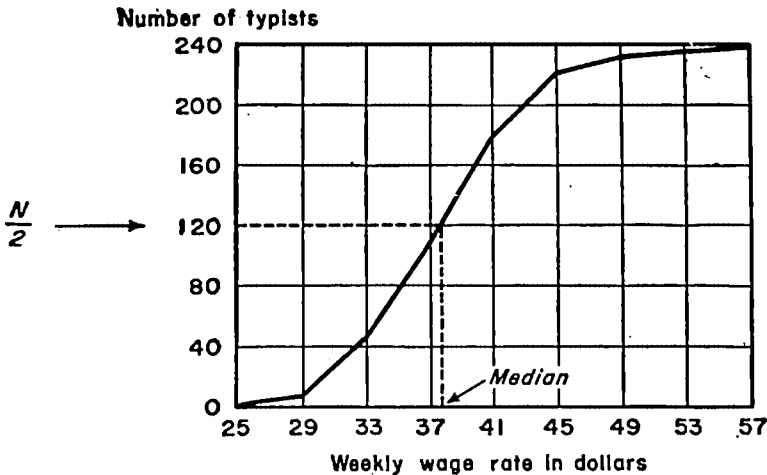


Chart 11.1. Graphic Analysis of Median through "Less than" Ogive.

Source: *Studies in Labor Statistics*, No. 2, National Industrial Conference Board, *Clerical Salary Survey of Rates Paid*, April 1949.

graph. In that case, the perpendicular is dropped from the point where the two ogives meet.

Of course, the median may be found graphically from a percentage ogive. Here very clearly $N/2$ is the point that has 50% of the items on each side of it.

Analogously, we can find the quartiles, the deciles, and the percentiles by means of graphic analysis. Then we no longer use $N/2$ but the appropriate fraction involving N , for instance $N/4$ for the first quartile.

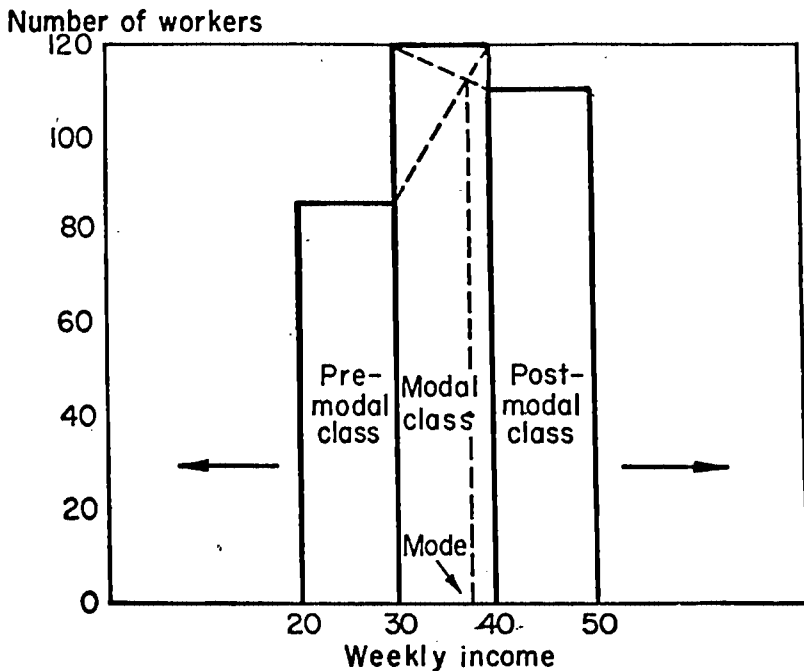


Chart 11.2. Graphic Analysis of Mode through Partial Histogram for Data in Table 11.4.

Source: Table 11.4.

Mode

On a histogram the modal class is, of course, the class with the tallest column. We can find the modal value within the modal class by a method which is the geometric counterpart

of the "difference" method that we used for algebraic computation. From the point where the top of the premodal rectangle borders on the modal rectangle, we first draw a line to the opposite corner of the modal rectangle. Then we draw a line from the point where the top of the postmodal rectangle meets the modal rectangle, to the opposite corner of the modal rectangle. Where these drawn lines meet, we drop a perpendicular to the X -axis. Where this perpendicular meets the X -axis, we can read the mode. This procedure is illustrated in Chart 11.2.

THE GEOMETRIC MEAN

Concept of the Geometric Mean

The arithmetic mean is a member—the most prominent member—of a group of averages which may be thought of as the "family" of means. Its other members are the geometric mean, the harmonic mean, and the quadratic mean. Of these three minor means the geometric mean is most important and the only one we need discuss here.

Like the arithmetic mean, the geometric mean is a computed measure and depends upon the size of each of the values in the series. But the geometric mean is not on the level of addition, sums and differences; rather it is on the level of multiplication, products and ratios. In short, it is not arithmetic; it is what its name implies, geometric. The difference between the arithmetic mean and the geometric mean has some similarity to the difference between an arithmetic grid and a semilogarithmic grid.

The geometric mean is the N th root of the *product* of N items. If there are two items, we take the square root; if three, the cube root; and so on. Since every item makes its presence felt in the geometric mean, this mean is affected by extreme values but not so much as is the arithmetic mean. The geometric mean is never larger than the arithmetic mean; on occasion it may turn out to be the same as the arithmetic mean, but usually it is smaller. If there are zeros or negative values in the series, the geometric mean cannot be used.

Computation of the Geometric Mean

The geometric mean of the series 2, 4, 8 is the cube root of $2 \times 4 \times 8$ or $\sqrt[3]{64} = 4$.

In ungrouped data, each item in the series is represented by the symbol X . The number of X 's is n . The geometric mean is represented by the symbol G , and the formula for finding the geometric mean is as follows:

$$G = \sqrt[n]{\text{product of the } X\text{'s.}}$$

If an ungrouped series has a number of items larger than two, or if a series is arranged in a frequency distribution, then the geometric mean is computed by means of logarithms. In fact, the geometric mean is sometimes referred to as the logarithmic mean. When logarithms are used, the geometric mean can be calculated like an arithmetic mean; that is, the logarithm of the geometric mean is equal to the arithmetic mean of the logarithms of the individual values. Thus, for ungrouped data, the formula is

$$\log G = \frac{\sum \log X}{N}$$

For grouped data, the formula is

$$\log G = \frac{\sum f \log X}{N}$$

Uses of the Geometric Mean

Of the uses for the geometric mean, three important examples are here given.

1. *Averaging ratios.* There are two plants in an industry and we wish to estimate the average percentage increase in production from 1955 to 1956. One plant has increased production 20%, the other 50%. Thus the 1956 production is 120% of 1955 production in the first plant and 150% in the second plant. If we want to give equal importance to each of them regardless of their absolute production figures, then we use

the geometric mean. The geometric mean comes to .134%. This means an average increase of 34%.*

The geometric mean is thus useful in averaging ratios, percentages, and rates. It is particularly important in economics and business statistics in index-number construction.

2. *Computing average rate of increase or decrease.* If a population increases 10% in ten years, the annual increase is *not* 10% $\div 10 = 1\%$, which would be the arithmetic mean. In such cases the concept of the geometric mean leads to a formula used to compute population changes and amounts at compound interest, and in analogous situations.

3. *Averaging a series approximating a geometric progression.* If production figures of firms in an industry are, in millions of tons, respectively 2, 4, 8, 16, 32, 64, 128, the geometric mean should be used. The computation of G is

$$G = \sqrt[7]{2 \times 4 \times 8 \times 16 \times 32 \times 64 \times 128};$$

or

$$\log G = \left(\frac{\log 2 + \log 4 + \log 8 + \log 16 + \log 32 + \log 64 + \log 128}{7} \right) = 1.2041$$

and

$$G = 16.$$

The arithmetic mean $\bar{X}$ of this series is 36.6; $\bar{X}$ is large compared with G since $\bar{X}$ is more affected by extreme values.

In practical work we usually do not find geometric progressions, but we do find series approximating them. In such situations the geometric mean is appropriate.

Limitations

Important limitations to the use of the geometric mean are that it is not at all well known, is not easy for the layman to grasp, and is relatively difficult to compute.

* On the other hand, if we wish to take account of the absolute production figures behind these percentages, then we use the weighted arithmetic mean (see pages 130-134).

Summary

1. The median is a "position" average. It is the *value* in a series such that one-half of the items in the series is below it in magnitude and one-half of the items in the series is above it in magnitude.

2. The median cannot be found unless there is an array or a frequency distribution. In ungrouped data, the median is found by counting the items and in grouped data by interpolation.

3. The median has a mathematical property,

$$\Sigma|x| = \text{a minimum,}$$

which comes into consideration in finding certain measures of variation. The fact that the median is a positional average makes it appropriate in certain specific situations.

4. Positional measures, resembling the median but not measures of central tendency, are the quartiles, the deciles, and the percentiles.

5. The mode is the value which occurs most frequently in a series.

6. If a mode can be established for ungrouped data, an array should be made in order to find it. For frequency distributions, in basic statistics, we find the "crude" mode, which is an approximation; the "true" mode cannot be found without advanced methods. The method for finding the crude mode, discussed here, is the "difference" method.

7. Unusually shaped distributions raise special problems regarding the usefulness of the mode.

8. The median and related positional measures may be found from an ogive, and the mode may be found from a histogram.

9. The geometric mean is a type of mean chiefly useful in averaging ratios, and in computing average rates of increase or decrease.

Measurement of Masses: Comparison of the Principal Averages

1. Location of the Three Averages within the Frequency Distribution

The *arithmetic mean* is like the center of gravity in the frequency distribution; it is a point of balance. If we consider the frequency polygon as cut out of metal or wood, then in order to hold it so that its base is horizontal we would have to hold it at a point on the horizontal which is the arithmetic mean. The frequency distribution may also be envisaged as a seesaw where an item at a large distance from the point of balance needs a number of items close to it to obtain balance; here again the mean is at the point of balance. Thus, the mean balances the *differences* from it. On the other hand, if we divide the frequency polygon at the *median* value, we get two *areas* of equal size; the median balances the number of items regardless of their value. And the *mode* is the point on the horizontal axis of the frequency polygon which is located exactly below the *peak* of the curve.

2. Comparison of the Location of the Three Averages

In a symmetrical curve, the mean, the median, and the mode are identical. That is, the value that consists of the sum of all

the items in the series divided by the number of the items is the same as the value that divides the series in half, and the same as the value that occurs most frequently. This identity is easily seen from a graph of a symmetrical distribution in Illustration 12.1.

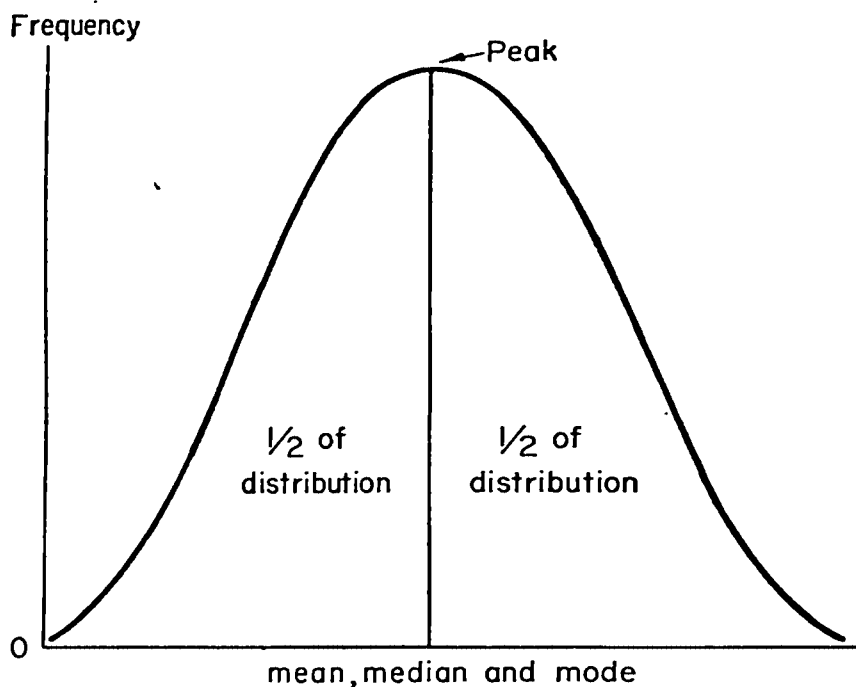


Illustration 12.1. Location of the Mean, the Median, and the Mode on a Symmetrical Curve.

If we were to add high values to the symmetrical series, the frequency distribution would be pulled to the right; that is, the distribution will be skewed to the right or positively skewed.* Since the mean is influenced by every value in the series, the new extreme high values added to the distribution pull the *mean* up (that is, to the right on the graph) most, whereas the *median*, which is a position average, is responsive only to the number of items which have been added, but not to their magnitude. There-

* See Chapter 9, pages 121-123.

fore, the median's upward (rightward) movement is not as great. The *mode*, however, is not responsive to the magnitude of the high values nor to the number of items added, and therefore will remain where it was. Hence, in a positively skewed distribution, the mean, median, and the mode have the respective positions shown in Illustration 12.2.

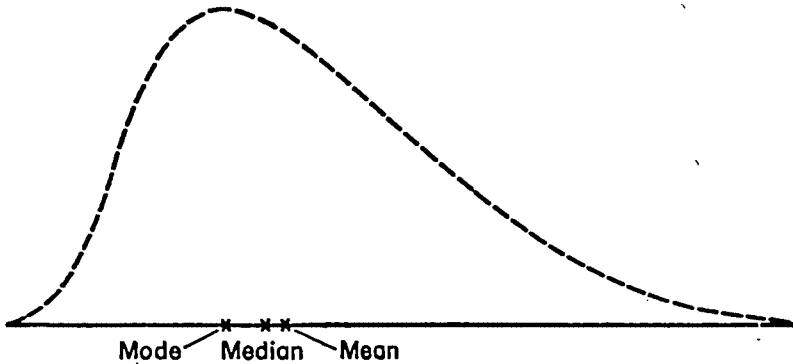


Illustration 12.2. Relative Positions of Mean, Median, and Mode in a Right-Skewed Distribution (Schematic).

On the other hand, in a left-skewed distribution, where the arithmetic mean is being pulled towards the low-valued items (that is, to the left), the positions of the averages are changed. There the *mean* is the smallest of the three values, the *median* is next larger, and the *mode* is the largest of the three, as shown in Illustration 12.3.

From the constancy of the sequence of these three measures in right-skewed and left-skewed distributions, and from his observation of a regularity in the relationship of their values to each other in moderately skewed distributions of continuous data, Karl Pearson arrived at a method for finding the mode. Starting with the distance between the mean and mode in moderately skewed distributions, Pearson found that the median divides this distance into two *unequal* parts, the larger of which is twice the size of the smaller. The larger is the distance between the mode and the median, and the smaller is the distance between the mean and the median. The distance between the mode

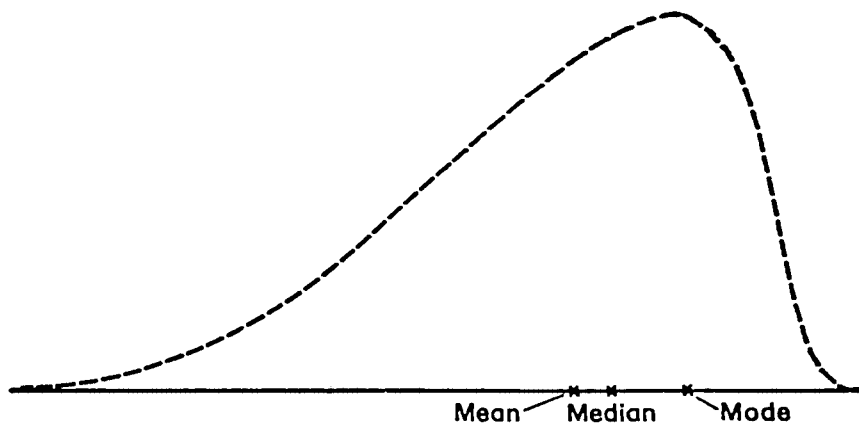


Illustration 12.3. Relative Positions of Mean, Median, and Mode in a Left-Skewed Distribution (Schematic).

and the median is twice the distance between the median and the mean. Thus, Pearson arrived at an empirical formula for the mode based upon the fact that we can express the mode in terms of the mean and the median. The mode is the mean minus three times the distance between the mean and the median. In Illustrations 12.4 and 12.5 this relationship is shown graphically.

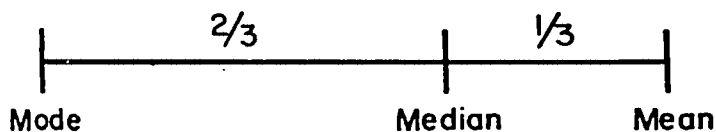


Illustration 12.4. Relative Positions of Mean, Median, and Mode in Moderately Right-Skewed Frequency Distribution.

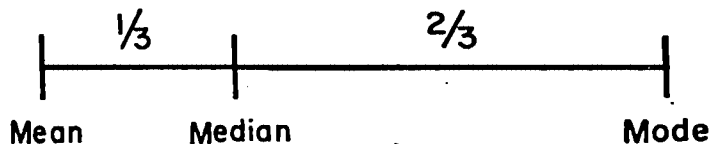


Illustration 12.5. Relative Positions of Mean, Median, and Mode in Moderately Left-Skewed Frequency Distribution.

Thus, Pearson's empirical formula for the mode is

$$\text{mode} = \text{mean} - 3(\text{mean} - \text{median}).$$

That is, the mode is equal to the mean minus three times the difference between the mean and the median.

The mode thus found by Pearson's empirical formula for a distribution is usually not identical with the mode found by the difference method (see pages 152-155).

3. Methods of Obtaining the Three Averages

The *arithmetic mean* is a computed average as opposed to the *median* and the *mode*, which are positional averages. This distinction, of course, does not mean that there is no calculation in finding the median and the mode for grouped data, but that the value of every item in the series goes into the making of the mean; it does not enter into the median and the mode in the same way. This basic difference results in certain other differences discussed below.

4. Effect of Extreme Values on the Three Averages

As one consequence of its being a computed average, the *arithmetic mean* is greatly affected by extreme values at either end of the series. Since the inclusion of one atypical item (either very small or very large) may cause considerable change in the mean, otherwise based on a number of typical items, the mean is in such a case not a representative description of a series.

In a given series, the *median* is not affected by the value of extreme items. But if items are added to one extreme end of the series whose median has already been established, the *number of items added will change the median but their value will not affect it*.

Extreme values do not influence the *mode* at all when it really measures central tendency, even if these values are added after this average has been established.

5. Effect of Open-End Classes on the Averages

An open-end class is a class lacking one limit. Hence it is impossible to determine the mid point of such a class, and it is therefore impossible to determine the *arithmetic mean*.

Since the *median* is based on locating a value in the middle of a distribution, lack of information at the ends of the distribution has no influence. In order to locate the median, we need to know the number of frequencies in each class up to the median class and the total frequency of the distribution, but not the class limits at the lower or upper ends of the distribution.

The *mode* is not affected by the presence of open-end classes, unless the modal class is itself open-ended; in this latter situation there is no central tendency, as in a J curve or a reverse-J curve.

6. Varying Class Intervals and the Averages

The *arithmetic mean* can be computed only by the long method when there are varying class intervals, but even then the result may differ widely from the mean for ungrouped data.

What varying class intervals usually indicate is that the distribution is heavily skewed. Errors caused by the mid-point assumption will offset each other completely in a symmetrical distribution, but in a distribution which is decidedly skewed, errors remain because of incomplete offsetting. Hence the mean for grouped data, in a distribution with varying class intervals, may differ *greatly* from the mean obtainable from the ungrouped data.

The *median* is not affected by varying class intervals since in finding the median we are concerned only with the size of the interval of the median class.

The modal class is the one with the highest frequency. Finding the *mode* thus entails a comparison of various class frequencies. But this comparison is possible in distributions with varying class intervals only on the basis of frequency densities. For example, the largest class frequency in a distribution may be due solely to the large interval of this class. The use of frequency

densities, however, bases the "crude" mode on an additional assumption involved in the concept of frequency densities.

7. Use of Averages in Further Computation

Another effect of the fact that the *mean* is a computed average is that it may be used in further computation, such as combining the means of two or more series. *Medians* and *modes* cannot be combined since they are values based on position in some particular series, and for two or more series combined, the median or modal position would have to be discovered anew.

The fact that the mean is a computed average is important for its use in advanced statistics, as against the use of the median and the mode.

8. Mathematical Properties

The *mean* has three leading mathematical properties:

1. The algebraic sum of the deviations of the items from the mean is equal to zero; that is, as we have put it in symbols, $\Sigma x = 0$.

2. The sum of the squared deviations of the items from the mean is smaller than the sum of the squared deviations of the items from any other value; that is, $\Sigma x^2 = \text{a minimum}$.

3. If we substitute the mean for each item in the series and then total these substitutions, the total thus obtained will be equal to the total of the original data; that is, $N\bar{X} = \Sigma X$.

The *median* has one mathematical property of importance: the sum of the deviations of the items from the median, *signs ignored*, is less than the sum of the deviations of the items from any other value, signs ignored, or $\Sigma |x| = \text{a minimum}$.

The *mode* has no simple mathematical properties.

9. Arrangement of Data and the Three Averages

The *mean* can be found from the raw data, the array, the frequency distribution, and indeed without individual items or classes if the sum of the items and their number are known.

The *median* cannot be found without an array or a frequency distribution.

The *mode* can be found from a frequency distribution, and sometimes from an array if there is noteworthy repetition of a value.

10. Obtaining the Averages from Graphs

The *median* may be found from an ogive, either a "less than" or an "or more" or both combined. The *mode* may be found from a histogram. The *mean* cannot be found graphically since it is a computed average.

APPROPRIATENESS OF THE THREE AVERAGES

There is no single answer to the question: Which average is to be used to describe statistical data?

There are situations where none of the three averages is fully satisfactory. For example, if the number of items in a series is very small, no one of these three types of average may have any import as a measure of central tendency. Moreover, if the items are what has been called "freakishly deployed" (as when concentrated at one end), an average for the series can be found but it is not descriptive of the series.

The data available may exclude the use of certain of the averages. For instance, in open-end distributions, the *mean* cannot be accurately found, but the *median* or the *mode* can be used. In a situation of bimodality, the mode makes no sense as a measure of central tendency.

The characteristics of the averages and the demands of the problem to be solved usually determine which one is to be used. The data may call for a certain average.

In such problems as the average score on personnel tests or the average productivity rate of workers, where the values in the series are really ranks, the *median* is the average to use. These scores and rates are not additive—that is, they indicate not unit

quantities but rather the position of an individual with respect to other individuals. Therefore, a positional average is appropriate, namely the median.

The *arithmetic mean* is the most commonly used and best known of the averages, and is preferred unless precluding circumstances are present, such as extreme values at either end of the series, or open-end classes, or varying class intervals, or unless we definitely wish to establish the most frequent value or some other positional average. Where further computational techniques are involved in the investigation, the mean is the average to be used.

Great care should be used in choosing an average adaptable to a series. On occasion, it may even be advisable to work out all three averages and present them, although, to be sure, this procedure creates an added burden for the reader as well as for the statistician. But the added burden is preferable to the use of a single average that may be an incomplete description.

But considering all these points, the statistician, guided by the desire to present an accurate picture of the data and to command respect, is the final judge of which average is most appropriate.

Summary

1. The mean, median, and mode are located at the same point in a symmetrical frequency distribution. In moderately asymmetrical series, the relative positions of the three averages lead to Pearson's empirical formula for the mode.

2. The mean is a computed average, whereas the median and the mode are positional.

3. Extreme values in the series affect the utility of the mean, but not of the median or the mode.

4. The presence of open-end classes excludes the use of the mean, but not of the median and sometimes not of the mode.

5. Varying class intervals usually make the mean unreliable, but do not affect the median. The mode may be found in such a case but only by reclassification through frequency densities.

6. Means may be combined, but not medians and modes.

7. The mean has three mathematical properties which make it indispensable in advanced statistical work. The median has one mathematical property, while the mode has none.

8. The median and the mode cannot be found unless there is an array or a frequency distribution, but the mean can be found also from raw data, and even sometimes without knowing the individual items.

9. The median and the mode may be found graphically, but not the mean.

10. What average should be used in a situation depends upon the sufficiency of the data, the distribution of the data, the classification system, and the problem to be solved.

CHAPTER

13

Measurement of Masses: Variation, Skewness, Kurtosis

VARIATION

Use of Measures of Variation or Dispersion

Each one of the averages is a typical figure; a measure of central tendency, but none of them tells us how the items in a series are distributed relative to this measure; that is, how the items scatter around the measure of central tendency. Two series may have the same average, but the items may scatter widely around the average in one case and be highly congregative in the other. Hence we must look for measures of scatter or dispersion around the average.

For example, in two cities the mean family income is the same, \$5000 a year; but in one we find on closer scrutiny a middle-class population with the overwhelming majority of annual incomes approximately \$5000, whereas in the other city we find a wide difference in income, with very poor and very rich people. Describing the income picture in these two cities through the mean alone would conceal the wide difference in their social-economic structure.

Thus we see that an average alone does not tell the full story. It is hardly fully representative of a mass unless we know the manner in which the individual items scatter around it. In short, a further description of the series is necessary if we are to gauge how representative the average is. The measurement of the scatter of the mass of figures in a series about the average is called a measure of variation or dispersion; it is this measure that informs us how representative the average is as a description of the mass. If the items scatter widely, then the average by itself is an attenuated representation. If the items scatter in a very narrow range, then the average would appear to be highly representative.

Using a measure of dispersion also permits us to compare two series with regard to their variability. The study of variation may also be looked upon as a means of determining uniformity. A high degree of variation means small uniformity; a low degree of variation means great uniformity.

What a measure of dispersion gives us, in conjunction with an average, is a description of the structure of the distribution and the role of the individual item in it.

A measure of dispersion can be used in its absolute form, or in a relative form for comparisons. We shall be concerned in basic statistics with four measures of absolute dispersion, and with three measures of relative dispersion. The measures of absolute dispersion are (1) the range; (2) the semi-interquartile range, also called the quartile deviation; (3) the average deviation; (4) the standard deviation. The three measures of relative dispersion, or "coefficients," are adaptations of measures of absolute dispersion.

ABSOLUTE VARIATION: POSITIONAL MEASURES

The Crude Range

This measure (usually referred to simply as the range) is the easiest measure of dispersion to compute and the easiest

to understand, but it is the least useful and informative. It is really no more than a rough measure of variation. The range is the difference between the highest and the lowest values in the series. Thus, it does not tell us anything about the distribution of values in the series relative to a typical value. Moreover, single extreme values at either end or both ends of a series can drive the range severely upward and distort understanding of the series itself. Thus the range, though easily calculated, is but a crude measure of dispersion.

But the range has some value in such statistical analyses as stock-exchange prices, where the low of a stock and the high are shown. The range is also used in statistical quality control as a ready measure of variation (see Appendix to Chapter 23).

In the arrayed data of wages of twenty typists on page 97, the range is the difference between the highest and the lowest values, or between \$44 and \$30; that is, the range is \$14. For grouped data as in Table 10.3 (see page 135), we take as the range the difference between the upper limit of the class of highest value and the lower limit of the class of lowest value. This is the difference between \$1 and \$15, or \$14.

In open-end distributions, the range cannot be ascertained unless information is available concerning the extreme values.

The Semi-Interquartile Range or the Quartile Deviation *

By eliminating the lowest 25% and the highest 25% of the items in a series, we are left with the central 50%, which are ordinarily free of extreme values. The total structure of the distribution is not shown by this measure, but a range is shown in a refined form.

These cut-off points which exclude the lower and the upper

* The term "quartile deviation," which is usually symbolized by QD , has been strongly criticized by some statisticians on the ground that it is not a measure of deviations from an average, but rather a distance on a scale (see p. 181). We give the term here as an alternative to "semi-interquartile range" since it has been so widely used. Similarly, on p. 202, in discussing relative variation measured in terms of the quartiles, we have used the term "coefficient of quartile deviation" although here too the above criticism holds.

25% of the items are the first and third quartiles. The term "quartile" is used to designate the values which divide the series into quarters. We have already discussed their general characteristics on page 151.

To obtain a measure of variation we use the distance between the first and the third quartile (the interquartile range), but customarily take half of this distance (the semi-interquartile range). This semi-interquartile range or quartile deviation may be looked upon as a type of range which is free from the distortions created by the presence of extreme values. It may be used in conjunction with the median.

The procedure for finding the semi-interquartile range or quartile deviation consists first of locating the first quartile, then of locating the third quartile, then subtracting the first quartile from the third quartile, and then dividing the result of this subtraction by two. In symbols, where Q stands for the semi-interquartile range,

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

The Quartiles for Ungrouped Data

For finding Q_1 and Q_3 , the basic logic is the same as for finding the median.

To estimate the first quartile, for ungrouped data, the method is as follows:

1. We start with the fraction $\frac{N + 1}{4}$.*
2. In the array, find the value corresponding to the $\frac{N + 1}{4}$ th item; this value is the first quartile.
3. It may be, however, that the $\frac{N + 1}{4}$ th item is a fraction. If that is the case, it is necessary to interpolate between the

* $N + 1$ is used in the numerator here in ungrouped data where we count items, whereas in grouped data where we do not know the items, we count spaces.

two values corresponding to the two items between which it falls.

Thus, if there are 36 items in a series, then $\frac{N+1}{4}$ is the 9.25th item. In this case, find the value of the 9th item in the array, and the value of the 10th item in the array, and then find the value of the 9.25th item by finding a value $\frac{1}{4}$ of the way between the 9th and 10th items. This last value is the first quartile.

This procedure is illustrated in Table 13.1 by finding the first quartile for the arrays of data in Tables 11.1 and 11.2.

TABLE 13.1. FINDING THE FIRST QUARTILE Q_1 AND THE THIRD QUARTILE Q_3 FOR THE ARRAYS OF DATA IN TABLES 11.1 AND 11.2.

	Q_1	Q_3
Data of Table 11.1; $N = 15$	$\frac{N+1}{4} = 4$ The 4th item is 25.50, $\therefore Q_1 = 25.50$	$\frac{3(N+1)}{4} = 12$ The 12th item is 32.50, $\therefore Q_3 = 32.50$
Data of Table 11.2; $N = 24$	$\frac{N+1}{4} = 6.25$ The 6th item is 50, the 7th item is 51, $\therefore Q_1 = 50.25$	$\frac{3(N+1)}{4} = 18.75$ The 18th item is 79, the 19th item is 82, $\therefore Q_3 = 81.25$

The procedure for estimating the third quartile for ungrouped data is basically the same. We start here with the fraction $\frac{3(N+1)}{4}$ and proceed as we did in finding the first quartile from ungrouped data. This procedure is also illustrated in Table 13.1 for the same data.

The Quartiles for Grouped Data

The method of finding the first and the third quartiles for a frequency distribution also follows the logic of the method for

finding the median (that is, the second quartile). The formula for finding the median in this case, it will be recalled, is

$$\text{median} = l_1 + \left(\frac{\frac{N}{2} - \Sigma f_1}{f_{\text{med}}} \right) i.$$

The formula for finding the first quartile for grouped data is therefore:

$$Q_1 = l_1 + \left(\frac{\frac{N}{4} - \Sigma f_1}{f_{Q_1}} \right) i,$$

where l_1 = the lower limit of the Q_1 class, the first class whose cumulative frequency equals or exceeds $N/4$ items;

Σf_1 = the sum of all the frequencies in the classes preceding the Q_1 class;

f_{Q_1} = the frequency in the Q_1 class; and

i = the size of the class interval in the Q_1 class.

The formula for finding the third quartile for grouped data is as follows:

$$Q_3 = l_1 + \left(\frac{\frac{3N}{4} - \Sigma f_1}{f_{Q_3}} \right) i$$

where l_1 = the lower limit of the Q_3 class, the first class whose cumulative frequency equals or exceeds $3N/4$ items;

Σf_1 = the sum of all the frequencies in the classes preceding the Q_3 class;

f_{Q_3} = the frequency in the Q_3 class; and

i = the size of the class interval in the Q_3 class.

The semi-interquartile range Q can now be found by substituting in the formula

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

In Table 13.2, the first and third quartiles are found for the frequency distribution in Table 11.3, as is Q .

Thus Q is a difference between values, whereas Q_1 and Q_3 are values.

Q and the Median

We have now explained and defined the semi-interquartile range Q , which is one measure of variation. Q cannot be compared directly with the average deviation and the standard deviation, the other two important measures of variation. But Q can be compared indirectly with these other measures, and the first step toward making such indirect comparison is to add Q to the median and to subtract Q from the median, thus establishing a range symbolized by median $\pm Q$.* Then we can establish the percentage of items falling within the range median $\pm Q$, in a normal distribution. Later in this chapter we shall establish the percentage of items (in a normal distribution) falling within the range

$$\text{average} \pm \text{average deviation}$$

and

$$\text{average} \pm \text{standard deviation}$$

Different percentages of items fall within these three ranges. These differences permit us to compare these three measures of variation.

The foregoing explains why we take Q with the median. An example of computing median $\pm Q$ is given in Table 13.2.

The relationship among the measures of variation, which is established precisely for a normal distribution, holds approximately for moderately asymmetrical distributions.

With the median, which is the second quartile, the other two quartiles give us three values with which to cut through the series. Between the value of the first quartile and the value of the third quartile, that is, between Q_1 and Q_3 , exactly 50%

* Read this expression as "median plus and minus Q ." In the statistical notation of this book the symbol $\pm$ means "plus and minus"; in algebra this same sign means "plus or minus" and the reader should fix the difference in mind to avoid possible confusion.

TABLE 13.2. FINDING Q_1 , Q_3 , Q , AND THE MEDIAN $\pm Q$ FOR THE FREQUENCY DISTRIBUTION IN TABLE 11.3.

Finding Q_1		Finding Q_3	
$N = 260$,	$\frac{N}{4} = 65$	$N = 260$,	$\frac{3N}{4} = 195$
$l_1 = 50$		$l_3 = 200$	
$\Sigma f_1 = 46$		$\Sigma f_3 = 187$	
$f_{q_1} = 59$		$f_{q_3} = 31$	
$i = 50$		$i = 100$	
$Q_1 = 50 + \left(\frac{65 - 46}{59}\right)50$		$Q_3 = 200 + \left(\frac{195 - 187}{31}\right)100$	
$= 50 + 16.1$		$= 200 + 25.8$	
$= 66.1 \text{ workers}$		$= 225.8 \text{ workers}$	

Exactly 50% of the industrial establishments in the city of Omega on July 1, 1956, have between 66.1 workers and 225.8 workers.

$$\begin{aligned} &\text{Finding } Q \text{ and median } \pm Q \\ Q &= \frac{225.8 - 66.1}{2} = \frac{159.7}{2} = 79.9; \\ &\text{median} = 127.8 \text{ workers,} \\ &\text{median} + Q = 207.7 \text{ workers,} \\ &\text{median} - Q = 47.9 \text{ workers.} \end{aligned}$$

Therefore *approximately* 50% of the establishments in the city of Omega on July 1, 1956, have between 47.9 workers and 207.7 workers.

of the values in the series fall. In a perfectly symmetrical distribution, it is clear that the first quartile and the third quartile are equidistant from the median, but symmetrical distributions are rare in actual life. In a right-skewed distribution, Q_3 is farther away from the median than Q_1 ; in a left-skewed distribution, Q_1 is farther away from the median than Q_3 .

In a symmetrical distribution, the median $\pm Q$ will bring us back exactly to the quartile values. Therefore, by definition, exactly 50% of the items will be found within this range. But in a moderately asymmetrical distribution, where the quartiles are not equidistant from the median, the median $\pm Q$ leads us to values close to, but not exactly at, the quartile values, and conse-

quently the number of items included in this range will only approximate 50%.*

Thus, in Table 13.2, we see that between the first quartile of 66.1 workers and the third quartile of 225.8 workers, we find *exactly* 50% of all establishments. But between 47.9 and 207.7 workers (median $\pm Q$) we find *approximately* 50% of the establishments.

Characteristics of Q

If the semi-interquartile range is very small, then it bespeaks small variation or large uniformity of the middle items. Thus, Q is of use in comparing variation or uniformity in different distributions. It is extremely valuable in measuring variation in open-end distributions.

It has been said that Q is not a measure of variation or dispersion since it really does not show the *scatter around an average*, but rather a distance on a scale. That is, Q is not itself *measured from an average*, but it is a positional measure. Consequently some statisticians speak of Q as a measure of *partition* rather than as a measure of deviation or dispersion or variation.

In order really to measure dispersion in the sense of scatter, we have to find the deviation of each item *from* an average, so that the deviation of each item makes itself felt. Deviation from an average is the concept involved in the two other measures of absolute dispersion, to which we now turn: the average deviation and the standard deviation.

ABSOLUTE VARIATION: COMPUTED MEASURES

Variety is not only the spice of life, but also a keystone of statistics.

The raw material out of which the computed measures of variation—the *average deviation* and the *standard deviation*—are

* If we cut the frequency polygon at the quartiles, we get exactly 50% of the area under the polygon. But if we cut the polygon at two other points, we are not sure to get 50% because the part of the polygon thus cut out is not of the same shape as when we cut it at the quartiles.

composed is the distance of each item in a series from a "norm." The norm here is a measure of central tendency. These concepts of type and of deviation from type are reflected in everyday speech. For example, the statement "He is very tall" means statistically that the given individual's height deviates widely from the average height. Likewise, the statement "Her wages are low" means statistically that the wage of the given individual deviates from the average wage.

The way we make use of the distances of the items from the average distinguishes the average deviation from the standard deviation.

THE AVERAGE DEVIATION

The average deviation (also called mean deviation) is the average distance of the items in a series from their average. That is, we find the deviation of each item from the arithmetic mean or median, and take the mean of these deviations. Every distance of an item is considered without regard to whether it is in the direction of more than or less than the average. In short, we disregard pluses and minuses.

Average Deviation for Ungrouped Data

The symbol for the average deviation is A.D. (some statisticians use M.D. to symbolize mean deviation). The symbol for the deviation of each item X from the arithmetic mean $\bar{X}$ of the series is small or lower-case x (that is, $X - \bar{X} = x$).^{*} Vertical lines around x , that is $|x|$, mean "disregarding signs." The x 's may be visualized as in Illustration 13.1. Thus, $\Sigma|x|$ denotes the total deviation in the series.

The formula for finding the *arithmetic mean of these deviations*, or the *average deviation*, for ungrouped data, is

$$\text{A.D.} = \frac{\Sigma|x|}{N}$$

^{*} This symbol x is also used if the deviation is taken from the median.

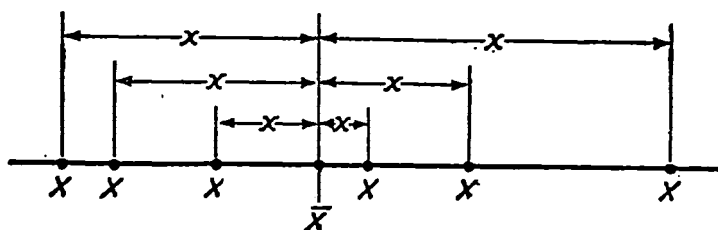


Illustration 13.1. Schematic Diagram Showing the Distance of Each Item from the Arithmetic Mean ($X - \bar{X} = x$), Ungrouped Data.

Signs must be disregarded because otherwise, if the mean of the series is being used as the value from which the deviations are taken, the sum of the deviations will be equal to zero, as we saw in the chapter on the arithmetic mean (page 141).^{*} The median is sometimes used as the average from which to find the deviations because a significant characteristic of the median is that the sum of the deviations of the items from the median is a minimum when signs are ignored. Despite this theoretical advantage of the median which makes the sum of the deviations more stable, the mean is more frequently used.

An example of the way in which the average deviation is found for ungrouped data is given in Table 13.3. The average deviation there is found to be \$8.80. This result means that on the average the rentals in this housing project differ by \$8.80 from the mean rental. The average deviation may be helpful in comparing the spread of rentals in another project or another city where the mean rental is approximately the same. Thus the extent of uniformity in rentals can be compared.

Average Deviation for Grouped Data

In a frequency distribution, the raw material of our computation—the distance of every item from the average—becomes the distance of the representatives of the items in each class, namely,

^{*} Some statisticians think that signs should be ignored for a different reason, namely, that every deviation has the same importance whether above or below the average. Hence a deviation of -3 or $+3$ is still a deviation of 3 .

TABLE 13.3. AVERAGE DEVIATION FOR MONTHLY RENTALS IN UNITS OF PUBLIC HOUSING DEVELOPMENT IN A LARGE METROPOLITAN AREA, MAY 1, 1959.

X	$ x $
\$30.00	\$20.00
54.00	4.00
42.00	8.00
69.00	19.00
58.00	8.00
37.00	13.00
48.00	2.00
53.00	3.00
60.00	10.00
49.00	1.00
\$500.00	\$88.00
$\Sigma X = \$500.00$	$\Sigma x = \$88.00$
$N = 10$	$N = 10$
$\therefore \bar{X} = \$50.00$	$\therefore \text{A.D.} = \$8.80$

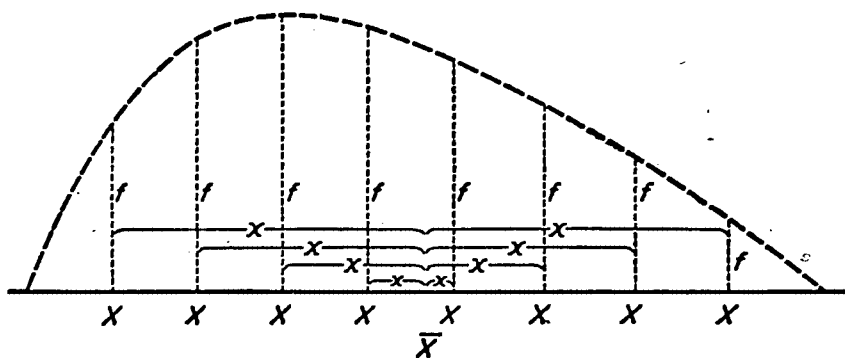


Illustration 13.2. Schematic Diagram Showing the Distances of the Mid Values from the Arithmetic Mean with Frequencies, Grouped Data.

the mid points, from the average. Of course, each mid point has to be taken as many times as there are items in the class which it represents. This procedure is shown schematically in Illustration 13.2.

For a frequency distribution, the average deviation may be found by the following method:

1. Find the deviation of the mid point of each class from the arithmetic mean or the median. Disregard the sign and symbolize this by $|x|$.*

2. In each class, multiply $|x|$ by the frequencies in the class. Symbolize by $f|x|$.

3. Add together $f|x|$ for each class, obtaining $\Sigma f|x|$.

4. Divide $\Sigma f|x|$ by N .

The formula for the average deviation for grouped data thus is

$$\text{A.D.} = \frac{\Sigma f|x|}{N}$$

Table 13.4 illustrates how the average deviation is found for the frequency distribution in Table 10.3. The average deviation

TABLE 13.4. COMPUTATION OF AVERAGE DEVIATION BY LONG METHOD FOR FREQUENCY DISTRIBUTION OF SALES CHECKS IN THE ALPHA STORE IN DALLAS, TEXAS, SEPTEMBER 1, 1956.

Class	Mid point, X	Class frequency, f	Deviation of mid point from mean, ¹ $X - \$7.76 = x $	Deviation times frequency, $f x $
\$1 and under \$3	\$2	3	\$5.76	\$ 17.28
\$3 and under \$5	\$4	9	3.76	33.84
\$5 and under \$7	\$6	25	1.76	44.00
\$7 and under \$9	\$8	35	.24	8.40
\$9 and under \$11	\$10	17	2.24	38.08
\$11 and under \$13	\$12	10	4.24	42.40
\$13 and under \$15	\$14	1	6.24	6.24
		100		190.24

¹ See Table 10.3.

$$N = \Sigma f = 100, \quad \Sigma f|x| = 190.24$$

$$\text{A.D.} = \frac{\Sigma f|x|}{N} = \frac{190.24}{100} = \$1.90$$

* The symbol here is the same as in ungrouped data, but refers to the distance of the mid point from the average, not the distance of the original item.

there is found to be \$1.90. This result means that on the average sales checks differ from the mean of \$7.76 by \$1.90. Thus, we may compare, for instance, this variation with the variation in sales checks on a day in the middle of the month in the same store when the mean sale was about the same. It may show that there is greater variation in sales on the first of the month than in the middle of the month. Thus, there would be less uniformity in sales on the first of the month than in the middle of the month.

There is a short method for finding the average deviation for a frequency distribution, but it is not used extensively in practical work and we shall not consider it here.*

A.D. and the Average

For the same reason that prompts us to take a range of the median $\pm Q$, we may now seek to establish the percentage of items falling in a range of the average $\pm$ A.D. In a normal distribution, within the range of values

$$\text{mean} \pm \text{A.D.},$$

which is the same as

$$\text{median} \pm \text{A.D.},$$

57.5% of the items in the series fall. If the distribution is moderately skewed, the percent of the items within the range of $\bar{X} \pm$ A.D. will *approximate* 57.5. Thus, if the average deviation is comparatively small, then more than half of the items in the series fall within a small range around the average. This concentration would mean compactness of the distribution.

THE STANDARD DEVIATION

The *standard deviation* may be looked upon as a special form of the average deviation, since it too is based on the deviations of the individual items in the series from an average. The

* The short method is adequately described in Robert E. Chaddock, *Principles and Methods of Statistics*, Houghton Mifflin Co., pp. 156-158.

standard deviation makes use of the same raw material—the deviations of the individual items from the average. In this case the average used is always the arithmetic mean; and the deviations are squared, thus avoiding the problem involved in disregarding signs.

The *mean of the squared deviations* is known as the *variance*. In advanced statistics, the variance has importance. But in basic statistics we use the standard deviation, which is the square root of the variance and is a measure on the level of the original units rather than on the level of their squares.

The standard deviation is linked with a property of the arithmetic mean, namely that the sum of the squares of the deviations of the items in the series from their arithmetic mean is a minimum. That is, the sum of the squares of the deviations of the items in the series from any other value must be greater than that from the arithmetic mean.

The standard deviation is the most important measure of variation, and variation is one of the pillars of statistics. Hence, the standard deviation is one of the most important statistical concepts.

Standard Deviation for Ungrouped Data: Long Method

The symbol for the standard deviation is the small or lower-case letter s . Frequently, the Greek letter σ is used; it is the lower-case sigma.

For ungrouped data the standard deviation, s , is found by taking the difference of each item in the series from the arithmetic mean ($\bar{x}$), squaring this difference (x^2), summing all the squared differences (Σx^2), dividing by the number of items (N), and then extracting the square root. Or in a formula:

$$s = \sqrt{\frac{\Sigma x^2}{N}}.$$

Finding the standard deviation for ungrouped data is illustrated

TABLE 13.5. STANDARD DEVIATION (LONG METHOD) FOR MONTHLY RENTALS IN UNITS OF PUBLIC HOUSING DEVELOPMENT IN A LARGE METROPOLITAN AREA, MAY 1, 1959.

X	$x.$	x^2
\$30	-20	400
54	+ 4	16
42	- 8	64
69	+19	361
58	+ 8	64
37	-13	169
48	- 2	4
53	+ 3	9
60	+10	100
49	- 1	1
		1188

$$\bar{X} = \$50,* \quad N = 10, \quad \Sigma x^2 = 1188$$

$$s = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{118.8} = \$10.90$$

* See Table 13.3.

in Table 13.5. The standard deviation here is \$10.90, and is interpreted analogously to the average deviation.

Standard Deviation for Ungrouped Data: Short Method

There is an easier method for computing the standard deviation for ungrouped data: we guess a mean and correct for the error. If we guess the mean as zero, then the deviation of each item from this guessed mean is the same as the value of the item itself. Thus, \$30 deviates from a guessed mean of zero by 30. Therefore, we work with deviations without having to compute them.

Since we guess a mean, we must correct for the error. We proceed by taking the sum of the squared deviations about the guessed mean. The sum we get will always be too high since

TABLE 13.6. STANDARD DEVIATION (SHORT METHOD) FOR MONTHLY RENTALS IN UNITS OF PUBLIC HOUSING DEVELOPMENT IN A LARGE METROPOLITAN AREA, MAY 1, 1959.

X	X^2
\$30	900
54	2,916
42	1,764
69	4,761
58	3,364
37	1,369
48	2,304
53	2,809
60	3,600
49	2,401
<u>500</u>	<u>26,188</u>

$$N = 10, \quad \Sigma X = 500, \quad \Sigma X^2 = 26,188$$

$$\begin{aligned}
 s &= \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2} = \sqrt{\frac{26,188}{10} - \left(\frac{500}{10}\right)^2} \\
 &= \sqrt{2,618.8 - 2,500} = \sqrt{118.8} \\
 &= \$10.90
 \end{aligned}$$

the corresponding sum taken from the correct mean is always a minimum. The sum of the squared deviations about any value other than the correct mean, even about zero, must be higher than that about the correct mean. Thus, the correction factor must always be subtracted from the sum of the squared deviations about the guessed mean. The correction factor is the square of the arithmetic mean. Hence, the formula is

$$s = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2} = \sqrt{\frac{\Sigma X^2}{N} - \bar{X}^2}$$

Table 13.6 shows the computation of the standard deviation by the short method for the ungrouped data in Table 13.5. The result is necessarily the same as the one obtained by the long method for ungrouped data.

Standard Deviation for Grouped Data: Long Method

For grouped data, by the long method, we take (as in finding the average deviation) the difference between each mid point and the arithmetic mean. But here we square this difference for each class, multiply the squared difference by the frequencies in the class, sum all these products, divide by the number of items in the series, and then extract the square root. Or in a formula:

$$s = \sqrt{\frac{\sum fx^2}{N}}$$

An example of finding the standard deviation for a frequency distribution by the long method is given in Table 13.7. Again the interpretation is analogous to that for the average deviation.

TABLE 13.7. COMPUTATION OF STANDARD DEVIATION BY LONG METHOD FOR FREQUENCY DISTRIBUTION OF SALES CHECKS IN THE ALPHA STORE IN DALLAS, TEXAS, SEPTEMBER 1, 1956.

Class	Mid point, X	Class frequency, f	Deviation of mid point from mean,* $X - \$7.76 = x$	Squared deviation, x^2	Squared deviation times frequency, fx^2
\$1 and under \$3	\$2	3	5.76	33.18	99.54
\$3 and under \$5	\$4	9	3.76	14.14	127.26
\$5 and under \$7	\$6	25	1.76	3.10	77.50
\$7 and under \$9	\$8	35	.24	0.06	2.10
\$9 and under \$11	\$10	17	2.24	5.02	85.34
\$11 and under \$13	\$12	10	4.24	17.98	179.80
\$13 and under \$15	\$14	1	6.24	38.94	38.94
		100			610.48

* See Table 10.3.

$$N = 100, \quad \sum fx^2 = 610.48$$

$$s = \sqrt{\frac{\sum fx^2}{N}} = \sqrt{\frac{610.48}{100}} = \sqrt{6.1048} = \$2.47$$

Standard Deviation for Grouped Data: Short Method

There are short methods for computing the standard deviation for grouped data. Like the short method for finding the arithmetic mean, one of the short methods for finding the standard deviation is based upon class-interval steps. In the short method, a mean may be guessed anywhere in the distribution, but it simplifies calculation to guess some mid point. The class whose mid point is guessed as a mean is 0 class interval away from itself; the preceding class, next lower in value, is -1 step interval away from this class; the next lower class is -2 ; and so on. The class following the class in which the mean has been guessed is $+1$ step interval away from this class; the next higher is $+2$ step intervals away; and so on.

The step-interval deviation, d , in each class is squared to get d^2 . The squared step-interval deviation for each class is taken as many times as there are frequencies. This gives us fd^2 for each class.

For each class we also multiply the nonsquared step interval by the frequency; giving fd for each class.

The column of fd^2 is summed. The sum of fd^2 is divided by the number of items, N . The column of fd is summed (algebraically, with due regard for signs); this sum is also divided by the number of items, and then the result of this division is squared.

$\Sigma fd^2/N$ is an expression based upon step deviations about a *guessed* mean; this expression must be larger than if taken about the actual mean since the sum of the squared deviations of the items from the actual mean is a minimum. Thus a correction factor will always have to be *subtracted* from $\Sigma fd^2/N$. This correction is analogous to that for the arithmetic mean by the short method, $\Sigma fd/N$, but here must be taken squared $(\Sigma fd/N)^2$. The entire expression must be reduced from the level of squares to the linear level. This reduction is accomplished by taking the square root of the expression.

Adjustment must be made for the size of the class interval since we have operated in terms of steps. The procedure is to multiply the square root by i , the size of the class interval.

Or in a formula.

$$s = i \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

Time and effort may be saved by manipulating this formula and obtaining the form

$$s = \frac{i}{N} \sqrt{N(\sum fd^2) - (\sum fd)^2}$$

The standard deviation for grouped data will be found worked out by the short method in Table 13.8. The short method must

TABLE 13.8. COMPUTATION OF STANDARD DEVIATION BY SHORT METHOD FOR FREQUENCY DISTRIBUTION OF SALES CHECKS IN THE ALPHA STORE IN DALLAS, TEXAS, SEPTEMBER 1, 1956.

Class	Mid point, X	Class frequency, f	Step deviation from class of guessed mean, d	Squared step deviation, d^2	Step deviation times frequency, fd	Squared step deviation times frequency, fd^2
\$1 and under \$3	\$2	3	-3	9	- 9	27
\$3 and under \$5	\$4	9	-2	4	-18	36
\$5 and under \$7	\$6	25	-1	1	-25	25
\$7 and under \$9	\$8 = $\bar{X}_d$	35	0	0	0	0
\$9 and under \$11	\$10	17	+1	1	+17	17
\$11 and under \$13	\$12	10	+2	4	+20	40
\$13 and under \$15	\$14	1	+3	9	+ 3	9
		100			-12	154

$$N = 100, \quad \sum fd = -12, \quad \sum fd^2 = 154$$

$$\begin{aligned}
 s &= \frac{i}{N} \sqrt{N(\sum fd^2) - (\sum fd)^2} = \frac{2}{100} \sqrt{100(154) - (-12)^2} \\
 &= \frac{2}{100} \sqrt{15,400 - 144} = \frac{2}{100} \sqrt{15,256} = \frac{2}{100} \times 123.5 = \frac{247}{100} \\
 &= \$2.47
 \end{aligned}$$

lead to the same result as the long method. Although the short-method formula is more involved, it makes calculation much easier since simpler figures are used.

THE NORMAL CURVE

One of the important uses of the normal curve in basic statistics is in connection with the standard deviation. This relation of the standard deviation to the normal curve has enhanced the utility of the standard deviation.

The normal curve is a special type of symmetrical frequency curve which is defined by a mathematical formula. In Illustration 13.3 a normal curve is shown. In basic statistics, we need

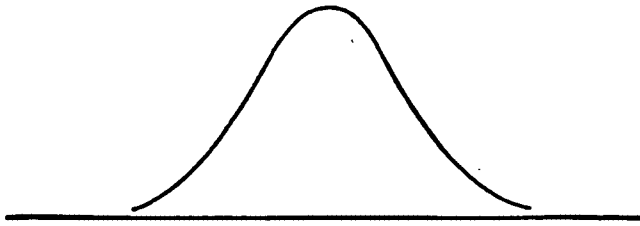


Illustration 13.3. Normal Curve.

not be concerned with the formula, but with the meaning of the concept of normal curve and its bearing on certain measures and aspects of statistical analysis. It is a tool indispensable in modern statistics.

What distinguishes the normal curve from other symmetrical curves? A normal curve is symmetrical not only with regard to skewness (as are all symmetrical curves), but it is also "symmetrical" with regard to peakedness since its peak is related to its dispersion by mathematical formula. Thus, normal curves constitute a family of curves whose peakedness is related to their dispersion in a strict way. Consequently the normal curve for any given trait is not the same as for some other trait. Normal curves have the same *general* shape but the normal curves of different distributions (for example, height or aptitude) differ in specific appearance.

The normal curve is a mathematical ideal, an abstraction or pure model. "Normal" here does not mean "what is to be expected or desired," nor "free from the abnormal," but is the technical name for this mathematically defined type of curve. The normal curve is not found in practical work, but it is an important concept for describing the form of distribution that would be taken by some continuous data in very large numbers, and toward which certain types of data in smaller numbers tend. We have already pointed out that statistics deals with approximations. Since certain types of data approximate the normal distribution, all laws of behavior of normal distributions may be applied to them.

What conditions make a normal distribution possible? The forces playing on each item in the distribution must be of the same importance, and no single force must be dominating. For instance, the distribution of heights of men is "normal" if everything that influences height is about equally operative throughout the mass and if each influence is independent of every other. The influences here are such causes as nutrition, climate, heredity. The use of a drug which retards growth or stimulates it would be a dominating influence which would prevent "normality." In natural phenomena, we find series (such as head and foot measurements) where these forces are numerous and independent; but in economics and business, such series will not be found (though many series approximate such conditions).

If a limit is imposed at one end of a distribution, a normal distribution is not possible. For example, if in aptitude testing for job placement we set a limit such that we do not employ any applicant whose grade falls below that limit, then the distribution of grades of the workers we do employ cannot be normal. Moreover, if we have a *heterogeneous* mass, we cannot expect a normal distribution. For example, if we have as customers Norwegians and Japanese, the shirt sizes we will stock cannot approximate "normality." Each group in itself may tend to normality, but they cannot do so when lumped together.

Normal curves are based upon variation and uniformity conjoined. That is, there is variation in every trait, but the variation is regularized. All heights differ, but within limits; nobody is 10 feet tall. Moreover, the pattern of variation is uniform and can be expressed mathematically; that is, there is as much variation below the mean as above the mean. Items differing from the mean (or median or mode—all of which are the same value in a normal curve) by the same amount in either direction occur with the same frequency. For example, if we have a very large number of measurements of the same object taken in milligrams, we find differences in the measurements. If we make a frequency distribution of these measurements, we will get a curve tending toward normality. At a given distance above and below the mean we will find the same number of measurements. In addition, the more a measurement deviates from the mean the less frequently it occurs.

Let us now return to the standard deviation and the analysis of the normal curve in terms of the standard deviation.

The Standard Deviation and the Normal Curve

In any normal curve, an exact percentage of items in the distribution falls within ranges established by the standard deviation in conjunction with the mean. If we cut through the distribution at one, two, and three standard deviations away from the mean, on both sides, we obtain areas of the distribution which contain certain percentages of all the items.

In a normal curve, between the range of the arithmetic mean plus 1 standard deviation and minus 1 standard deviation ($\bar{X} \pm 1s$), 68.27% of the items in the distribution fall; between the range of the arithmetic mean plus and minus 2 standard deviations ($\bar{X} \pm 2s$), 95.45% of the items fall; and within the range of the arithmetic mean plus and minus 3 standard deviations ($\bar{X} \pm 3s$), 99.73% of the items fall. Thus, in a normal curve, 6 times the standard deviation constitutes practically the whole range of the values in the entire series, as is shown

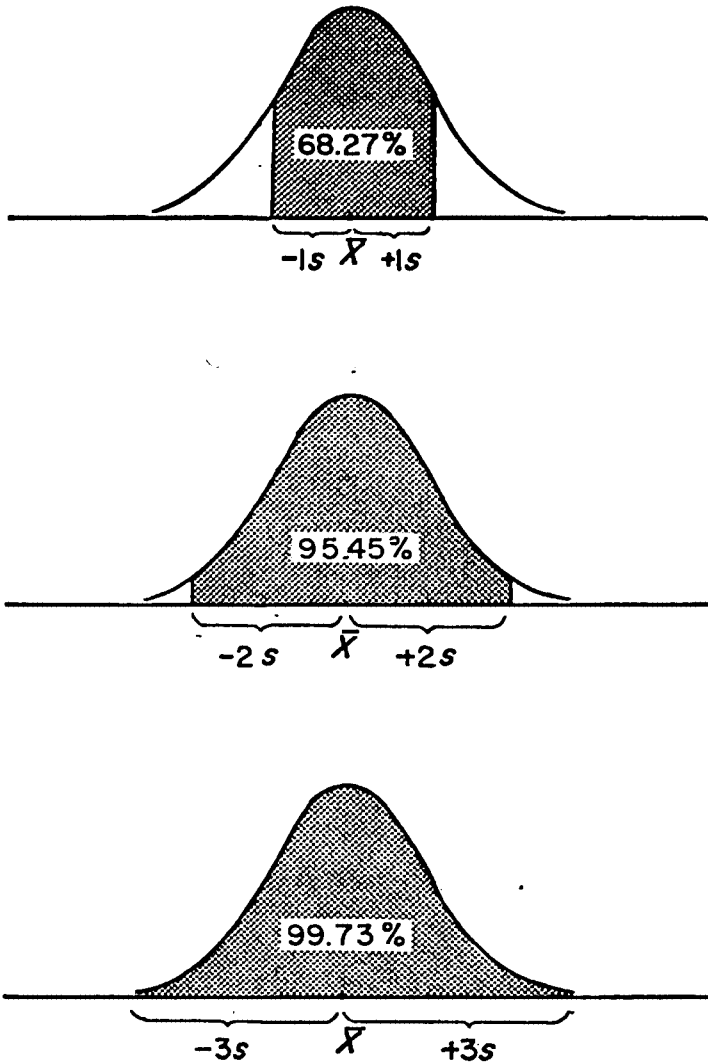


Illustration 13.4. Proportion of Items Included within $\pm 1s$, $\pm 2s$, $\pm 3s$, of the Arithmetic Mean in a Normal Distribution, Graphically Shown as Percentages of the Area under a Normal Curve. The area under the normal curve is never 100%, since the normal curve approaches the axis at both ends but never meets it.

graphically in Illustration 13.4. For moderately skewed distributions, the above percentages hold approximately.

This rigorous relation of the standard deviation to the normal curve permits us to put the normal curve to three uses of vast importance in practical work. These uses are:

1. From the normal curve we can find the number of cases *at* any given distance from the mean;
2. From the normal curve we can find the number of cases *lying between* any values in the distribution;
3. From the normal curve we can establish the probability that a case selected at random will lie above or below a given point on the scale.

Detailed instructions on how to use the normal curve in these three ways are contained in an Appendix at the end of this chapter.

Use of the Standard Deviation

Of all the measures of absolute dispersion, the standard deviation is most important. It is a keystone in sampling and in correlation.

In basic statistics, what uses can we make of the standard deviation * as a measure of variation? (1) We can compare the extent of variation or degree of uniformity by comparing the standard deviations of two or more series. (But valid comparisons of standard deviations can be made only when the arithmetic means of the series are about the same and the series are expressed in the same units.) Thus if a merchant receives from two different manufacturers sample shipments of the same commodity, whose essential variable quality is hardness, he will place his order with the manufacturer whose sample shipment has the smaller standard deviation. The desire for great uniformity (or small variation) in hardness leads this merchant to place greater

* These uses hold for other measures of dispersion as well as for the standard deviation. They are discussed here because of the practical importance of the standard deviation.

confidence in the manufacturer whose sample shipment shows a smaller standard deviation. (2) The standard deviation helps us to locate the place of an item in the distribution. If a salesman sells \$100 more than the mean sales of all salesmen in his organization, we cannot say how good a salesman he is unless we know where he falls in the distribution. If we know, however, that his sales volume is 3 standard deviations above the mean, we know that he is the best, or one of the best, salesmen. (3) We can use the standard deviation to gauge the representativeness of the mean. If we have two or more comparable distributions with the same mean, then it is in the distribution with the smallest standard deviation that the mean is most representative.

COMPARISON OF MEASURES OF ABSOLUTE VARIATION

Additional insight into the three important measures of absolute variation (Q , A.D., s) may be gained by comparing their characteristics, their advantages, and their disadvantages in six respects as follows:

1. Type of Measure

Q is a measure of location or partition, while A.D. and s are computed measures of dispersion into whose making the value of every deviation goes.

2. Relation to Averages

Q is not a measure of variation about a particular average, but a distance on a scale—a type of range—and is found without reference to an average. A.D. measures variation about the mean or the median, whereas s measures variation about the mean only.

3. Effect of Extreme Values

Q involves only the central half of the distribution and is therefore not affected by extreme values. A.D. is affected by

extreme values but less so than s , for s places great emphasis upon extremes since deviations are squared. For example, if the deviations x are 1, 2, 3, 10, then

$$\text{A.D.} = \frac{1 + 2 + 3 + 10}{4} = 4,$$

whereas

$$s = \sqrt{\frac{1 + 4 + 9 + 100}{4}} = 5.3.$$

4. Relation of Measures of Absolute Variation in a Normal Curve

A basis for comparison of the measures of dispersion can be advantageously established only through the concept of normal distribution where the percentage of items falling within a certain range can be definitively put. In a normal curve, the median $\pm Q$ encompasses 50% of the items in the series, while $\bar{X}$ (or the median, which is in the normal curve identical with $\bar{X}$) $\pm$ A.D. encompasses 57.5% of the items in the series.

But only the standard deviation is used to analyze the *whole* distribution, approximately 68% of the items falling within the range of $\bar{X} \pm 1s$, approximately 95% within the range of $\bar{X} \pm 2s$, and almost 100% within the range of $\bar{X} \pm 3s$.

Thus in a given series the semi-interquartile range Q is always the smallest of the three measures, and the standard deviation s is always the largest. In a normal distribution Q is about two-thirds as large as s , and A.D. is about four-fifths as large as s .

In practical work, one-sixth of the range is sometimes used as a rough estimate of the standard deviation since in a normal distribution the range equals approximately six standard deviations.

5. Relation of Measures of Absolute Variation to Algebraic Properties of the Averages

The sum of the deviations of the items from the median is a minimum when signs are ignored, and thus A.D. when taken

from the median is theoretically preferable. The mean alone is the basis for finding the standard deviation since the sum of the squared deviations of the items from the mean is a minimum. Since Q is not taken from an average, it has no relation to the algebraic properties of an average.

6. Extent of Use

Q must be used to measure variation in open-end distributions, or when it is desirable to rule out the influence of extreme values. The use of the average deviation is overshadowed by the use of the standard deviation. In correlation, in sampling, and in other aspects of advanced statistics; the standard deviation has become indispensable because of inherent properties.

RELATIVE VARIATION

There are situations where absolute measures of variation are not comparable. For instance, if a plant whose mean wage is \$30 a week has a standard deviation of \$10 and if a plant whose mean wage is \$100 a week also has a standard deviation of \$10, then this identical standard deviation of \$10 has very different meanings in these two plants. The same absolute amount of dispersion around \$30 and around \$100 offers no basis of comparison in itself. A range of from \$20 to \$40 (that is, $\$30 \pm 1s$) is comparatively important; a range of from \$90 to \$110 (that is, $\$100 \pm 1s$) is comparatively slight.

If it were used to compare these two series, the standard deviation by itself would lead us astray. To get a sound basis for comparison we have to consider the standard deviation relative to the magnitude around which the dispersion is measured.

Let us take another illustration. We may wish to know whether a group of workers show the same uniformity in their productivity that they show in their scores on an aptitude test. Comparison of the standard deviations by themselves would not make sense

here, since the standard deviation of productivity is in number of units produced whereas the standard deviation of scores on the aptitude test is in percent.

Thus, a measure of absolute dispersion does not suffice for comparing the variation of one series with another series where the averages are different for the same variable, such as wages; neither does it suffice for comparing the variation of one series (such as productivity rates) with that of another series measuring a different variable (such as scores on an aptitude test). No matter which measure of absolute variation is used, the measure is not sufficient for such comparisons.

Since the disturbing influence in comparing the computed measures of variation in two or more series is the presence of averages of different sizes or different units of measurement, we can achieve comparability only by disposing of the disturbing influence. We dispose of it by dividing each measure of dispersion by its respective average.* The result of this is called a measure of *relative variation*.

There are three measures of relative variation. The one most often used consists of a ratio of the standard deviation to the mean. This ratio bears the symbol V , and is called the *coefficient of variation*. It was developed by Karl Pearson. The formula for it is

$$V = \frac{s}{\bar{X}}$$

Thus, in our instance of wages in the two plants, the first plant has an s of \$10 and an $\bar{X}$ of \$30; therefore, $V = \$10/\$30$ or 0.33. This can also be written as 33%. The second plant has an s of \$10 also, but an $\bar{X}$ of \$100; therefore V in this case is $\$10/\$100$ or 0.1 or 10%. Hence, variation is relatively much larger in Plant I than in Plant II. This comparison can be made only through the use of measures of relative variation, not through the use of measures of absolute variation.

* But not when the quartile deviation (semi-interquartile range) is involved, since it is not taken around an average.

The second measure of relative variation is referred to as the *coefficient of average deviation* and is symbolized by $V_{A.D.}$. It consists of a ratio of the average deviation to the mean or the median (depending on which was used to find A.D.). Its formula is either

$$V_{A.D.} = \frac{A.D.}{\bar{A}} \quad \text{or} \quad V_{A.D.} = \frac{A.D.}{\text{median}}$$

The third measure of relative variation is referred to as the *coefficient of quartile deviation*. It bears the symbol V_Q , and is most appropriate in open-end distributions and distributions with extreme values. It consists of a ratio between the difference between the quartiles and the sum of the quartiles. Its formula is

$$V_Q = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Each one of these coefficients may be expressed in decimal form or percentage form.

In comparing variation in two or more distributions where measures of relative variation are called for, the same type of measure must be used. For example, V in one distribution is not to be compared with $V_{A.D.}$ or V_Q in other distributions.

Measures of relative variation introduce no new statistical concepts, but are rather adaptations of measures of absolute variation to particular situations in which comparison of variation could otherwise not be made.

MEASUREMENT OF SKEWNESS

Masses of unorganized, raw data give us no basis for comparing series. If we want to compare or describe masses of data, we need expressions that bring out characteristic features of these masses. Up to now we have considered central tendency and variability.

But there are two other comparable characteristics: asymmetry and peakedness; these are more specifically called skewness and kurtosis, respectively.

Pearson's Measure of Skewness

Measures of skewness have been developed. These measures tell us the direction and the extent of skewness. It will be recalled that in a symmetrical distribution the mean, median, and mode are identical. The more the mean moves away from the mode, the larger the asymmetry or skewness. The distance between the mean and the mode is Karl Pearson's basis for measuring skewness. According to Pearson's empirical formula (see page 167), the distance between the mean and the mode in moderately skewed distributions is equivalent to three times the distance between the mean and the median.

Skewness is most often measured in order to compare series. But the same amount of skewness has different meanings in distributions with small variation and in distributions with large variation. In order to make valid comparisons between the skewness in two or more distributions, we have to eliminate the disturbing influence of variation; such elimination is accomplished by dividing. To get the Pearsonian *coefficient of skewness* we divide the difference between the mean and the mode by the standard deviation. Thus the formula for the Pearsonian coefficient of skewness is

$$Sk_P = \frac{\text{mean} - \text{mode}}{\text{standard deviation}}$$

Since, as we have seen, in moderately skewed distributions

$$\text{mode} = \text{mean} - 3(\text{mean} - \text{median}),$$

we may substitute in the formula for skewness above, and remove the mode from the formula, as follows:

$$Sk_P = \frac{\text{mean} - [\text{mean} - 3(\text{mean} - \text{median})]}{\text{standard deviation}}$$

This becomes

$$Sk_P = \frac{\text{mean} - \text{mean} + 3(\text{mean} - \text{median})}{\text{standard deviation}};$$

or in a final form:

$$Sk_P = \frac{3(\bar{X} - \text{median})}{s}.$$

This Pearsonian coefficient of skewness is based on the fact that skewness in a perfectly symmetrical distribution will be zero; that is the mean, the median, and the mode will all be the same. Where the curve of the distribution is skewed to the right, the measure of skewness for a distribution will be positive since the mean will be greater than the median. Where the curve of the distribution is skewed to the left, the measure of skewness will be negative since the median will be greater than the mean.

The basis for measuring skewness by Pearson's method is schematically shown in Illustration 13.5.

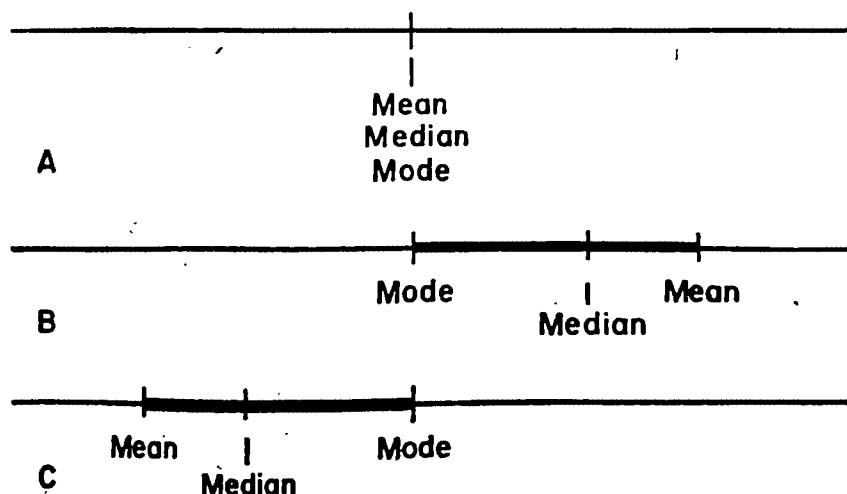


Illustration 13.5. Schematic Diagram Showing Basis of the Pearsonian Measure of Skewness. A, Symmetrical Distribution; B, Right-Skewed Distribution; C, Left-Skewed Distribution.

Let us find the Pearsonian measure of skewness for the frequency distribution of sales checks in the Alpha Store in Dallas, Texas, on September 1, 1956. In Tables 10.3 and 10.4, $\bar{X}$ was found to be \$7.76. In Tables 13.6 and 13.7, s was found to be \$2.47. The median for this distribution is \$7.74. Therefore, the Pearsonian coefficient of skewness is

$$\frac{3(7.76 - 7.74)}{2.47} = \frac{3(.02)}{2.47} = \frac{.06}{2.47} = +.024;$$

this means that the distribution is skewed to the right (positive skewness as shown by the plus sign), but very slightly.

But measures of skewness are used mainly for making comparisons between two or more distributions. As a description of one distribution alone, the interpretation of a measure of skewness is necessarily vague, as "slight skewness," "marked skewness," or "moderate skewness."

In our illustration, the measure of skewness, $+.024$, is useful in comparison with the direction and extent of asymmetry in sales-check data of another store. For instance, if the distribution of the sales checks in the Beta store (which is part of the same chain of stores) shows a Pearsonian coefficient of skewness of $-.432$, this coefficient indicates that Beta's skewness is to the left and much larger than Alpha's. Comparatively, this means that the distribution of Beta's sales checks is pulled towards the low-value checks, while Alpha's differs little from a symmetrical distribution around the mean sale.

The maximum amount of skewness from this Pearsonian formula is $+3$ or -3 , but skewness of more than $+1$ or -1 is rarely found.

Bowley's Measure of Skewness

It is also possible to measure skewness in terms of the quartiles. In a symmetrical distribution, the first and third quartiles are equidistant from the median. In asymmetrical distributions, the quartiles are not equidistant from the median, with the first quartile being farther away from the median than the third

quartile in the case of left skewness, and vice versa in the case of right skewness.

Since there is no difference between the distances of the quartiles from the median in a symmetrical distribution, any difference in their distances from the median is a possible basis for measuring skewness. An illustration of skewness measured by this approach is given in Illustration 13.6.

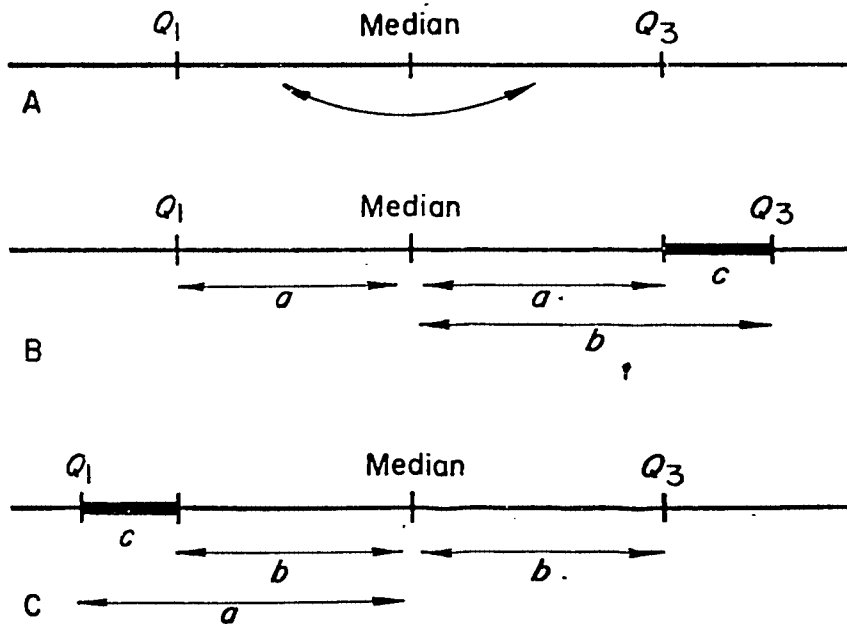


Illustration 13.6. Schematic Diagram Showing Basis of the Bowley Measure of Skewness. A, Symmetrical Distribution; B, Right-Skewed Distribution (a = distance median $- Q_1$, b = distance Q_3 - median, c = difference between distances a and b); C, Left-Skewed Distribution.

Just as we removed the influence of variation in finding the Pearsonian measure of skewness, so also we must remove this influence when using a quartile measure of variation. This removal is accomplished by using the full interquartile range* in

* The semi-interquartile range is sometimes used, and is equally acceptable logically.

what has come to be known as the Bowley measure of skewness after Arthur L. Bowley, who developed it.

The Bowley measure of skewness has the formula

$$Sk_B = \frac{(Q_3 - \text{median}) - (\text{median} - Q_1)}{\text{interquartile range}}$$

or

$$Sk_B = \frac{Q_3 + Q_1 - 2 \times \text{median}}{Q_3 - Q_1}$$

Bowley's measure of skewness has a maximum value of +1 or -1. In this measure .1 may be considered "moderate skewness" and .3 "marked skewness."

Wherever positional measures are called for, skewness should be measured by the Bowley method; thus this method is useful in open-end distributions and where extreme values are present.

In Table 13.2, Q_1 was found to be 66.1 workers and Q_3 , 225.8 workers. The median is 127.8 workers. Therefore the Bowley measure of skewness of this distribution is

$$\frac{(225.8 + 66.1) - 2(127.8)}{225.8 - 66.1} = +0.23.$$

The use of this measure is analogous to the use of the Pearsonian measure of skewness, but the two measures are not comparable to each other.

KURTOSIS

A fourth characteristic used for description and comparison of frequency distributions is the peakedness of the distribution. Measures of peakedness are known as measures of kurtosis. The computational aspects of kurtosis are beyond the scope of our discussion. The concept, however, can be understood at this point.

Kurtosis in Greek means "bulginess." In statistics, kurtosis

refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. The degree of kurtosis of a distribution is measured relative to the peakedness of a normal curve.

From the standpoint of kurtosis the normal curve is *mesokurtic*, which means "of intermediate peakedness." Flat-topped curves, on the other hand, are called *platykurtic*, while pronouncedly peaked curves are called *leptokurtic*. These three types are shown in Illustration 13.7.

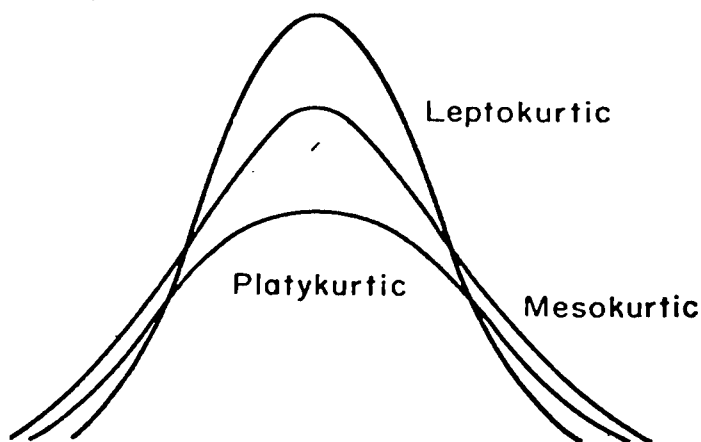


Illustration 13.7. Curve Types as to Kurtosis.
Leptokurtic, Mesokurtic, Platykurtic.

Summary

1. Measures of variation supplement averages in describing series, and show how representative the average is. Variation can be measured on an absolute basis and on a relative basis.
2. There are two positional measures of absolute variation: the range and the semi-interquartile range. The range is a rough measure of variation. The semi-interquartile range rules out

extreme values. The quartiles on which the semi-interquartile range is based are found analogously to the median.

3. The computed measures of absolute variation are the average deviation and the standard deviation. Both are based on the deviation of each item in a series from an average. In the average or mean deviation we take the mean of the deviations regardless of sign, but in the standard deviation we take the square root of the mean of the squared deviations.

4. Of the measures of absolute variation, the standard deviation is most important. The normal curve is analyzed in terms of the standard deviation, and these findings can be applied to series tending towards normality. The concept of the normal curve introduced in this chapter is more fully developed in the Appendix at the end of this chapter.

5. We can compare standard deviations in two or more series, we can locate items in a series through the standard deviation, and we can use it to gauge the representativeness of the mean.

6. Comparison of Q , A.D., and s may be made in terms of: type of measure, relation to averages, effect of extreme values, the normal curve, algebraic properties of averages, and extent of use.

7. Measures of relative variation are used to compare variation in series which differ in magnitude of their averages or in the units in which they are expressed.

8. The three coefficients of relative variation are the Pearsonian coefficient of variation, the coefficient of average deviation, the coefficient of quartile deviation.

9. Measures of skewness tell us the direction and extent of asymmetry in a series, and permit us to compare two or more series with regard to these. The two measures of skewness are the Pearsonian coefficient of skewness and the Bowley quartile measure of skewness. .

10. Kurtosis refers to the peakedness of a frequency curve. Measures of kurtosis use the peakedness of the normal curve as a reference.

APPENDIX—THE NORMAL CURVE

As Chapter 13 has explained, analysis of the normal curve in terms of the standard deviation permits certain practical uses of this normal curve. In the cases of distributions approximating normality, the relations discussed below hold approximately.

Ordinates and the Normal Curve

In a normal distribution, we can find the number of items located at any given distance from the mean. We do not draw or read a graph of a normal curve, but instead we use a table to find the information. The number of items located at any given distance from the mean corresponds to an ordinate at a given X value; this ordinate, of course, is measured along the vertical Y -axis. How do we arrive at the value of the ordinate from a table?

To answer the question we must know the number of items falling at the mean. Then we express the given value as its deviation from the mean and transform this deviation into standard-deviation units. Thus, if $\bar{X}$ is \$100 and s is \$10 and we wish to find the number of items falling at \$90, we find the deviation (ignoring signs) of 90 from 100 (which deviation is symbolized by x as we know from Chapter 13). This deviation of \$10, divided by s (in this case also \$10), is equal to 1.00. This deviation in standardized form is symbolized by x/s . From Table A13.1, Ordinates of the Normal Curve, we find that where the distance from the mean in standardized form, x/s , is 1.00, the ordinate is .6065 or 60.65% of the mean ordinate, that is to say, of the ordinate at the mean. See Illustration A13.1. In other words, the number of items falling at \$90 in this illustrative distribution is 60.65% of the number of items falling at the mean. If we know that there are 1000 cases falling at the mean of \$100, then there are approximately 607 cases at the value of \$90.

If we wish to find the number of items falling at \$112.50 in this

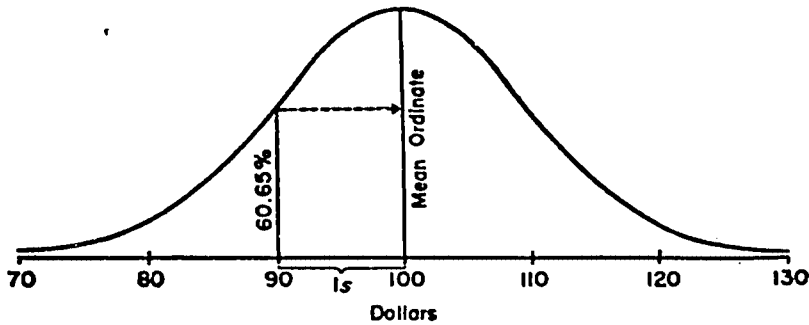


Illustration A13.1

illustrative distribution, we first find that this value is \$12.50 away from the mean of \$100. Since $s = \$10$, \$12.50 is 1.25 standard-deviation units away from the mean. From Table A13.1, we see that at 1.25 standard-deviation units from the mean there are .4578 or 45.78% of the items found at the mean ordinate. See Illustration A13.2. This per-

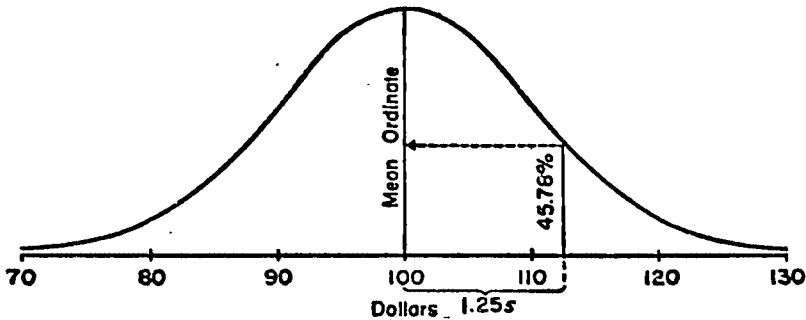


Illustration A13.2

centage was found by looking in the x/s column under 1.2 and then following the 1.2 line to the column headed .05 and reading off under that column at that line the figure .4578. Since we know there are 1000 cases at the mean, then we can conclude that there are approximately 458 cases at \$112.50.

If several ordinates in a normal distribution are found in the above manner, a rough draft of the appropriate normal curve can be obtained. Thus by this procedure we can compare visually a given distribution with a normal distribution having the same number of items if we know $\bar{X}$, s , and the number of items at $\bar{X}$. This comparison permits us

TABLE A13.1. ORDINATES OF THE NORMAL CURVE AS FRACTIONS OF THE ORDINATE AT THE MEAN

$\frac{x-s}{s}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	1.0000	1.0000	.9998	.9996	.9992	.9988	.9982	.9976	.9968	.9960
0.1	.9950	.9940	.9928	.9916	.9903	.9888	.9873	.9857	.9839	.9821
0.2	.9802	.9782	.9761	.9739	.9716	.9692	.9668	.9642	.9616	.9588
0.3	.9560	.9531	.9501	.9470	.9438	.9406	.9373	.9338	.9303	.9268
0.4	.9231	.9194	.9156	.9117	.9077	.9037	.8996	.8954	.8912	.8869
0.5	.8825	.8781	.8735	.8690	.8643	.8596	.8549	.8501	.8452	.8403
0.6	.8353	.8302	.8251	.8200	.8148	.8096	.8043	.7990	.7936	.7882
0.7	.7827	.7772	.7717	.7661	.7605	.7548	.7492	.7435	.7377	.7319
0.8	.7262	.7203	.7145	.7086	.7027	.6968	.6909	.6849	.6790	.6730
0.9	.6670	.6610	.6550	.6489	.6429	.6368	.6308	.6247	.6187	.6126
1.0	.6065	.6005	.5944	.5883	.5823	.5762	.5702	.5641	.5581	.5521
1.1	.5461	.5401	.5341	.5281	.5222	.5162	.5103	.5044	.4985	.4926
1.2	.4868	.4809	.4751	.4693	.4636	.4578	.4521	.4464	.4408	.4352
1.3	.4296	.4240	.4185	.4129	.4075	.4020	.3966	.3912	.3859	.3806
1.4	.3753	.3701	.3649	.3597	.3546	.3495	.3445	.3394	.3345	.3295
1.5	.3247	.3198	.3150	.3102	.3055	.3008	.2962	.2916	.2870	.2825
1.6	.2780	.2736	.2692	.2649	.2606	.2563	.2521	.2480	.2439	.2398
1.7	.2358	.2318	.2278	.2239	.2201	.2163	.2125	.2088	.2051	.2015
1.8	.1979	.1944	.1909	.1874	.1840	.1806	.1773	.1740	.1708	.1676
1.9	.1645	.1614	.1583	.1553	.1523	.1494	.1465	.1436	.1408	.1381
2.0	.1353	.1327	.1300	.1274	.1248	.1223	.1198	.1174	.1150	.1126
2.1	.1103	.1080	.1057	.1035	.1013	.0991	.0970	.0950	.0929	.0909
2.2	.0889	.0870	.0851	.0832	.0814	.0796	.0778	.0760	.0743	.0727
2.3	.0710	.0694	.0678	.0662	.0647	.0632	.0617	.0603	.0589	.0575
2.4	.0561	.0548	.0535	.0522	.0510	.0497	.0485	.0473	.0462	.0451
2.5	.0439	.0429	.0418	.0407	.0397	.0387	.0378	.0368	.0358	.0349
2.6	.0341	.0332	.0323	.0315	.0307	.0299	.0291	.0283	.0276	.0268
2.7	.0261	.0254	.0247	.0241	.0234	.0228	.0222	.0216	.0210	.0204
2.8	.0198	.0193	.0188	.0182	.0177	.0172	.0167	.0163	.0158	.0154
2.9	.0149	.0145	.0141	.0137	.0133	.0129	.0125	.0122	.0118	.0115
3.0	.0111									

Source: Data taken from *Tables for Statisticians and Biometricians*, edited by Karl Pearson, Cambridge University Press, London.

TABLE A13.2. AREAS UNDER THE NORMAL CURVE—PER CENT OF
TOTAL AREA

$\frac{x-s}{s}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.00	0.40	0.80	1.20	1.60	1.99	2.39	2.79	3.19	3.59
0.1	3.98	4.38	4.78	5.17	5.57	5.96	6.36	6.75	7.14	7.53
0.2	7.93	8.32	8.71	9.10	9.48	9.87	10.26	10.64	11.03	11.41
0.3	11.79	12.17	12.55	12.93	13.31	13.68	14.06	14.43	14.80	15.17
0.4	15.54	15.91	16.28	16.64	17.00	17.36	17.72	18.08	18.44	18.79
0.5	19.15	19.50	19.85	20.19	20.54	20.88	21.23	21.57	21.90	22.24
0.6	22.57	22.91	23.24	23.57	23.89	24.22	24.54	24.86	25.17	25.49
0.7	25.80	26.11	26.42	26.73	27.04	27.34	27.64	27.94	28.23	28.52
0.8	28.81	29.10	29.39	29.67	29.95	30.23	30.51	30.78	31.06	31.33
0.9	31.59	31.86	32.12	32.38	32.64	32.90	33.15	33.40	33.65	33.89
1.0	34.13	34.38	34.61	34.85	35.08	35.31	35.54	35.77	35.99	36.21
1.1	36.43	36.65	36.86	37.08	37.29	37.49	37.70	37.90	38.10	38.30
1.2	38.49	38.69	38.88	39.07	39.25	39.44	39.62	39.80	39.97	40.15
1.3	40.32	40.49	40.66	40.82	40.99	41.15	41.31	41.47	41.62	41.77
1.4	41.92	42.07	42.22	42.36	42.51	42.65	42.79	42.92	43.06	43.19
1.5	43.32	43.45	43.57	43.70	43.83	43.94	44.06	44.18	44.29	44.41
1.6	44.52	44.63	44.74	44.84	44.95	45.05	45.15	45.25	45.35	45.45
1.7	45.54	45.64	45.73	45.82	45.91	45.99	46.08	46.16	46.25	46.33
1.8	46.41	46.49	46.56	46.64	46.71	46.78	46.86	46.93	46.99	47.06
1.9	47.13	47.19	47.26	47.32	47.38	47.44	47.50	47.56	47.61	47.67
2.0	47.72	47.78	47.83	47.88	47.93	47.98	48.03	48.08	48.12	48.17
2.1	48.21	48.26	48.30	48.34	48.38	48.42	48.46	48.50	48.54	48.57
2.2	48.61	48.64	48.68	48.71	48.75	48.78	48.81	48.84	48.87	48.90
2.3	48.93	48.96	48.98	49.01	49.04	49.06	49.09	49.11	49.13	49.16
2.4	49.18	49.20	49.22	49.25	49.27	49.29	49.31	49.32	49.34	49.36
2.5	49.38	49.40	49.41	49.43	49.45	49.46	49.48	49.49	49.51	49.52
2.6	49.53	49.55	49.56	49.57	49.59	49.60	49.61	49.62	49.63	49.64
2.7	49.65	49.66	49.67	49.68	49.69	49.70	49.71	49.72	49.73	49.74
2.8	49.74	49.75	49.76	49.77	49.77	49.78	49.79	49.79	49.80	49.81
2.9	49.81	49.82	49.82	49.83	49.84	49.84	49.85	49.85	49.86	49.86
3.0	49.87									

Source: Data taken from *Tables for Statisticians and Biometricians*, edited by Karl Pearson, Cambridge University Press, London.

to judge how closely the given distribution approximates a normal distribution.

Areas under the Normal Curve

If we want to say what *proportion* of items in a normal distribution falls above a given value, below a given value, or between any two values, then the problem becomes one of areas rather than ordinates.

We already know (page 195) that in a normal frequency curve 68.27% of the items fall between $\bar{X} + 1s$ and $\bar{X} - 1s$, that 95.45% of the items fall between $\bar{X} + 2s$ and $\bar{X} - 2s$, and that 99.73% of the items fall between $\bar{X} + 3s$ and $\bar{X} - 3s$. Since the normal frequency curve is absolutely symmetrical (there are as many items at a given distance above $\bar{X}$ as there are at the same distance below $\bar{X}$), we know also that between $\bar{X}$ and $\bar{X} + 1s$, 34.13% of the items (that is, $68.27/2$) fall, and the same percentage between $\bar{X}$ and $\bar{X} - 1s$. Between $\bar{X}$ and $\bar{X} + 2s$, 47.72% (that is, $95.45/2$) of the items fall, and the same percentage between $\bar{X}$ and $\bar{X} - 2s$. Between $\bar{X}$ and $\bar{X} + 3s$, 49.87% (that is, $99.73/2$) of the items fall, and the same percentage between $\bar{X}$ and $\bar{X} - 3s$.

From Table A13.2 we can determine the area under a normal curve between the mean and any other value in a distribution. Thus "area" refers not only to a section of the distribution but also to a proportion of the total number of items in the distribution.

How are these areas established? We know certain areas under the curve: we know the area between the mean and values that are one, two, and three standard deviations away from the mean, and intermediate standard-deviation values. Therefore, we begin by expressing the deviation of a given value from the mean in terms of the standard deviation, that is, in standardized form, for example, 1.35s. From the table mentioned above we can read off the percentage of items that lie between the mean and the value expressed in standardized form—in other words, the area under the curve bordered by perpendiculars erected at the mean and at the given point. This is 41.15%. See Illustration A13.3.

We can put Table A13.2 to several important uses. Let us take as an illustration a normal distribution of grades * in an aptitude test given to 10,000 workers in a large manufacturing corporation. The mean

* This distribution of grades can be treated as continuous data since differences between successive grades are very small.

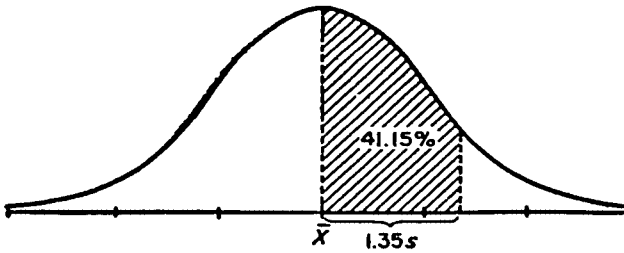


Illustration A13.3

grade on this test for these workers is 500. The standard deviation is 100.

Example 1A. We may wish to find the number of workers with grades above 375. We proceed as follows:

- How far is 375 from the mean? *Answer:* $500 - 375 = 125$.
- What is the standardized form of this difference of 125? *Answer:* $125/100 = 1.25s$.
- In a normal distribution what percentage of workers have a grade *higher* than the mean grade? *Answer:* 50%.
- What is the percentage of workers whose grades fall between the mean and a value 1.25s below the mean? *Answer* (read from Table A13.2 in the row 1.2 and the column .05): 39.44%.
- What, then, is the percentage of workers with a grade higher than 375? *Answer:* $39.44 + 50 = 89.44\%$.
- 89.44% of 10,000 workers is 8944 workers whose grades are higher than 375.

Illustration A13.4 portrays the situation just analyzed.

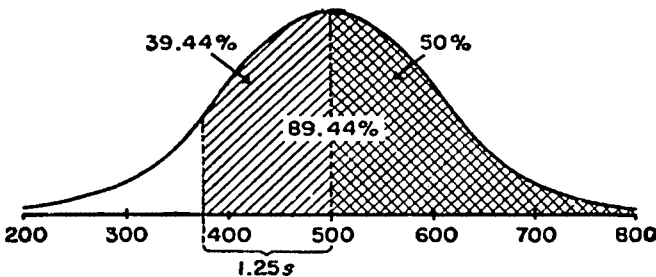


Illustration A13.4

Looked at another way the conclusion in "f" indicates that, if we chose a worker at random from among the 10,000 workers, the chances are about 90 out of 100 that his grade would be above 375. This type of interpretation can also be made in analogous situations.

Example IB. How many workers from this illustrative distribution have grades above 675?

a. $675 - 500 = 175.$

b. $175/s = 175/100 = 1.75s.$

c. In Table A13.2 we find that 45.99% of the workers' grades fall between $1.75s$ and $\bar{X}$.

d. $50\% - 45.99\% = 4.01\%.$

e. Thus 4.01 % of 10,000 = 401 workers have grades above 675. This situation is portrayed in Illustration A13.5.

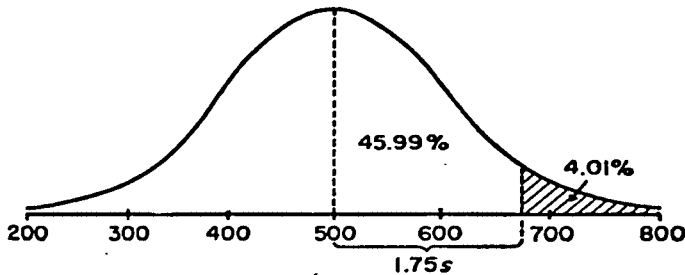


Illustration A13.5

Example IC. Another example: How many workers have grades below 350?

a. $500 - 350 = 150.$

b. $150/s = 150/100 = 1.5s.$

c. In Table A13.2 we find that 43.32% of the workers' grades fall between $1.5s$ and $\bar{X}$.

d. $50\% - 43.32\% = 6.68\%.$

e. Thus 6.68% of 10,000 = 668 workers have grades below 350. This situation is portrayed in Illustration A13.6.

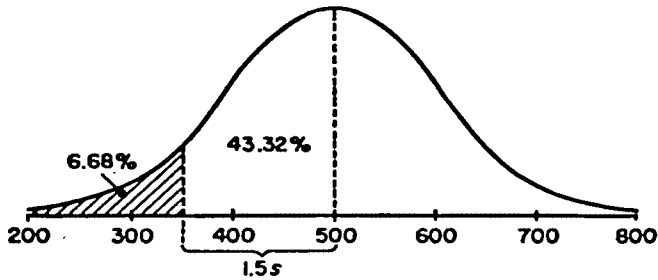


Illustration A13.6

Example 2A. We can find the percentage and number of items which fall between any two values in the distribution. For example, how many workers' grades fall between 450 and 525 in this distribution, where $\bar{X} = 500$, $s = 100$, and $N = 10,000$?

- a. $500 - 450 = 50$; $50/100 = .50s$ below the mean.
 $525 - 500 = 25$; $25/100 = .25s$ above the mean.

b. Between $\bar{X}$ and 450 (wherefore $x/s = .50s$), 19.15% of the grades fall, as can be seen in Table A13.2. Between $\bar{X}$ and 525 (wherefore $x/s = .25s$), 9.87% of the grades fall, as can be seen in Table A13.2.

c. Between 450 and 525, then, $19.15\% + 9.87\% = 29.02\%$ of the grades fall; thus 29.02% of the grades lie between 450 and 525.

d. Thus 29.02% of 10,000 = 2902 workers have grades between 450 and 525.

Illustration A13.7 portrays this situation.

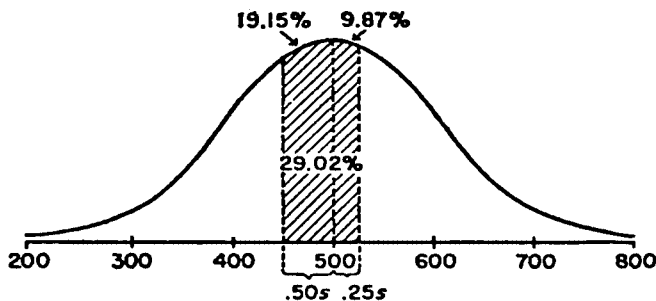


Illustration A13.7

Example 2B. How many workers have grades between 650 and 750 in this illustrative distribution? (This time both grades are higher than the mean grade.)

- a. $750 - 500 = 250$; $250/100 = 2.5s$ above the mean.
 $650 - 500 = 150$; $150/100 = 1.5s$ above the mean.

b. Between $\bar{X}$ and 750 (wherefore $x/s = 2.5s$), 49.38% of the workers' grades fall as found in Table A13.2. Between $\bar{X}$ and 650 (wherefore $x/s = 1.5s$), 43.32% of the workers' grades fall as found in Table A13.2.

c. Between 650 and 750, then, $49.38\% - 43.32\%$ of the workers' grades fall; or 6.06% of the workers fall between 650 and 750 in this distribution.

d. Thus 6.06% of 10,000 = 606 workers have grades between 650 and 750.

See Illustration A13.8 for a portrayal of this situation.!

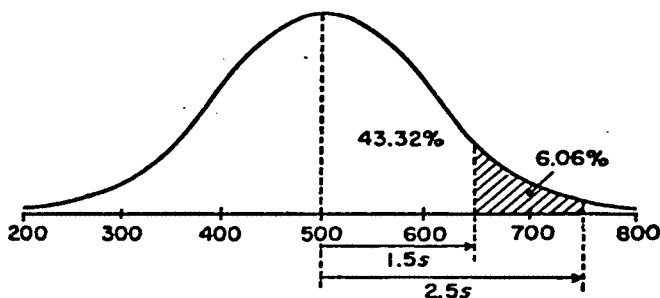


Illustration A13.8

Example 3A. We can find the *grade* above or below which a given percentage or number of workers' grades fall. For example, what is the grade above which the top 15% of the workers' grades fall in this illustrative distribution?

a. Between the mean item and the highest item in a normal distribution 50% of the items fall. The grade that marks off the upper 15% of the workers' grades must also mark off 35% of the grades between it and the mean grade.

b. We look for the figure closest to 35% in the body of Table A13.2. This figure is 35.08% and is 1.04s from $\bar{X}$. Since $s = 100$, $1.04s = 104$.

c. Therefore, the point we are looking for is 104 units above 500 or is 604.

d. Since 50% of the workers' grades fall above the mean and since 35% of the grades fall between the mean and 604, it follows that 604 is the grade above which 15% of the grades fall, or otherwise stated, the top 15% of the distribution of 10,000 workers have grades above 604.

This situation is portrayed in Illustration A13.9.

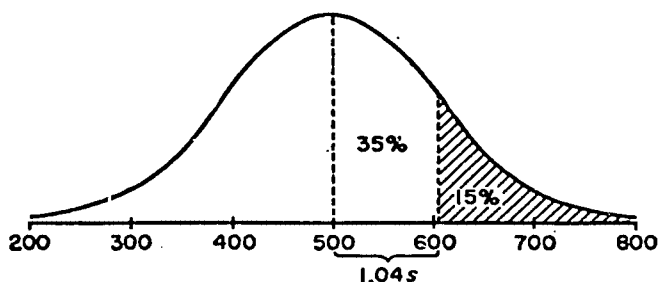


Illustration A13.9

Example 3B. What is the grade below which the bottom 10% of the workers' grades fall?

a. The grade that marks off the lower 10% of the workers must also mark off 40% of the workers' grades between it and the mean grade.

b. From Table A13.2 we find that between $1.28s$ and $\bar{X}$, 39.97% or approximately 40% of the workers' grades fall. Since $s = 100$, $1.28s = 128$.

c. The difference $500 - 128 = 372$. Between 372 and 500, 40% of the grades fall.

d. Since 50% of the workers have grades below the mean grade and 40% of the workers' grades fall between $\bar{X}$ and 372, therefore 10% of the grades fall below 372; or otherwise stated, the bottom 10% of this distribution of 10,000 workers have grades below 372.

This situation is portrayed in Illustration A13.10.

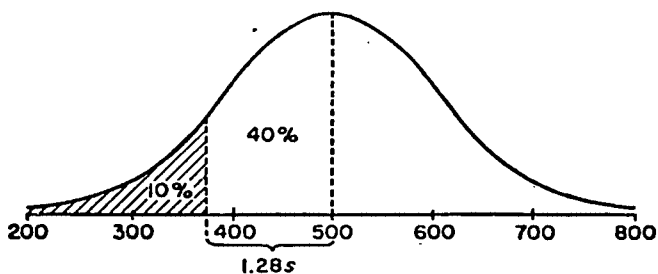


Illustration A13.10

Example 4. We can find the grades on either side of the mean between which a percentage of the grades falls. Between what grades does the middle 10% of the grades fall? To find the middle 10% of the grades we must find two grades between each of which and $\bar{X}$ 5% of the grades fall.

a. We look within Table A13.2 for the figure nearest 5%. This figure is 5.17 and is $0.13s$ from $\bar{X}$.

b. Since $s = 100$, $0.13s = 13$.

c. Computing $\bar{X} \pm 13$ gives $500 + 13 = 513$ and $500 - 13 = 487$.

d. Therefore the middle 10% of the workers' grades fall between 487 and 513.

Illustration A13.11 portrays this situation.

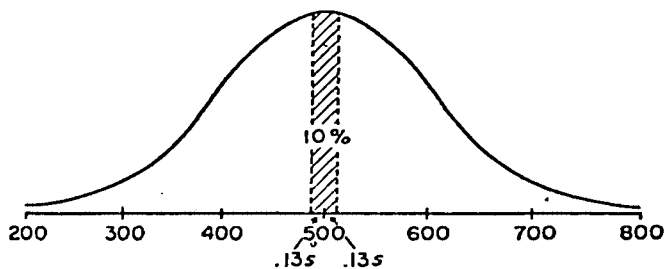


Illustration A13.11

In using the normal curve in practical work in the ways described in this Appendix we must be sure that the distribution approximates normality or tends towards normality. It must be remembered that "normal" does *not* mean the general type of distribution to be expected. Rather "normal" means the type that frequency distributions of *certain* variables tend toward when there is a large number of items. If there is good reason to suppose that the distribution of a given variable would approach normality if there were a greater number of items, then the principles that hold for a normal curve can prove of great value in analyzing the distribution.

The use of the principles underlying normal distributions is indispensable in sampling theory and its applications, as becomes clear from Chapters 21, 22, and 23 of this book.

CHAPTER

14

Introduction to Time-Series Analysis

What Is Time-Series Analysis?

In studying frequency distributions one of our main interests is in the variation of the items about the measure or measures of central tendency. How great, for example, is the variation in a distribution of wages, or prices, or sales observed at some particular point in time? But wages, prices, and sales also vary from one time period to another; these chronological variations will be the object of our study in time-series analysis.

Variation demands comparison. In time-series analysis, current data in a series may be compared with past data in the same series, for example a series of employment figures. We may also compare the development of two or more series over time; production figures of one firm in an industry for ten years may be compared with production figures of competitors and with the figures for the industry as a whole. These comparisons may afford important guide lines for the individual firm.

From comparison of past data with current data, we may seek to establish what developments may be expected in the future. Looking into the future through time-series analysis is called statistical forecasting.

Time-series analysis is of particular importance for the statistician who applies statistics to the fields of economics and business. Typical problems are the development of steel production over a number of years, the fluctuation of department-store sales within a year, and changes in agricultural employment or in commodity prices.

An economy is dynamic, not static; its dynamism is tied to the time factor. The idea of chronological movement is basic in economic analysis. Production, sales, and other types of economic data move through time, and we want to analyze them in motion—we wish to take a moving picture rather than a snapshot. If there were no variation, all sales, all incomes, all production data would move unchanged through time.

ELEMENTS OF TIME SERIES

If there were no variation in a time series, a graph of the data plotted over time would be a straight horizontal line. But there is variation, and when we plot time-series data on a graph we get “ups” and “downs.” What explains these “ups” and “downs”? A composite force is at work, that has pulled and pushed until the straight horizontal line that would have resulted from lack of variation assumes the up-and-down shape.

What are the components of the force? There are four: (1) trend; (2) seasonal variation; (3) cyclical variation; (4) irregular variation. Changes in data over a period of time are considered as the resultant of the combined impact of these four components.

Any series chronologically classified is in its raw form called original data. The original data and the four components of which they are the resultant are related by the equation

$$O = T \times S \times C \times I,$$

where O = original data,

T = trend,

S = seasonal variation,

C = cyclical variation, and
I = irregular variation.

In this chapter we shall deal in an introductory way with each of these four elements.

Trend

Trend, also called secular or long-term trend, is the basic tendency of production, sales, income, employment, or the like to grow or decline over a period of time. The concept of trend does not include short-range oscillations but rather steady movements over a long time.

What causes this growth or decline? In economic time series, growth in population is a main cause. The presence of more people means that more food, clothing, housing are necessary. Technological changes, discovery and exhaustion of natural resources, mass-production methods, improvements in business organization, and government intervention in the economy are other major causes for the growth or decline of many economic time series. In some cases, growth in one series involves decline in another; for example, the displacement of silk by rayon.

Electric-energy production is a good illustration of trend; it grows through the years. (See Chart 14.1.) Decline over the long term is also trend. The number of horses and mules on our farms has shown such a tendency in recent years. (See Chart 14.2.)

Infrequently, we come upon a time series over the long term which shows neither a tendency to grow nor a tendency to decline. An example of this is the population of Fall River, Massachusetts, from 1900 to 1950. (See Chart 14.3.)

Occasionally, structural changes take place in parts of our economy. The two World Wars and the Great Depression caused such far-reaching changes in several economic series that their development after the wars or after the depression

Electric energy in
billions of kilowatt hours

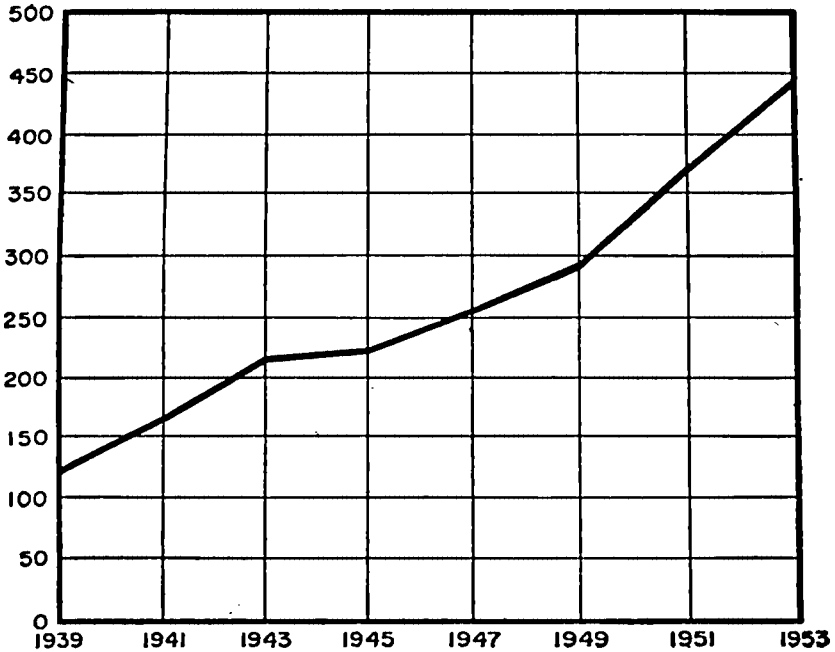


Chart 14.1. Production of Electric Energy in the United States, 1939-1953.

Source: Federal Power Commission, 1954.

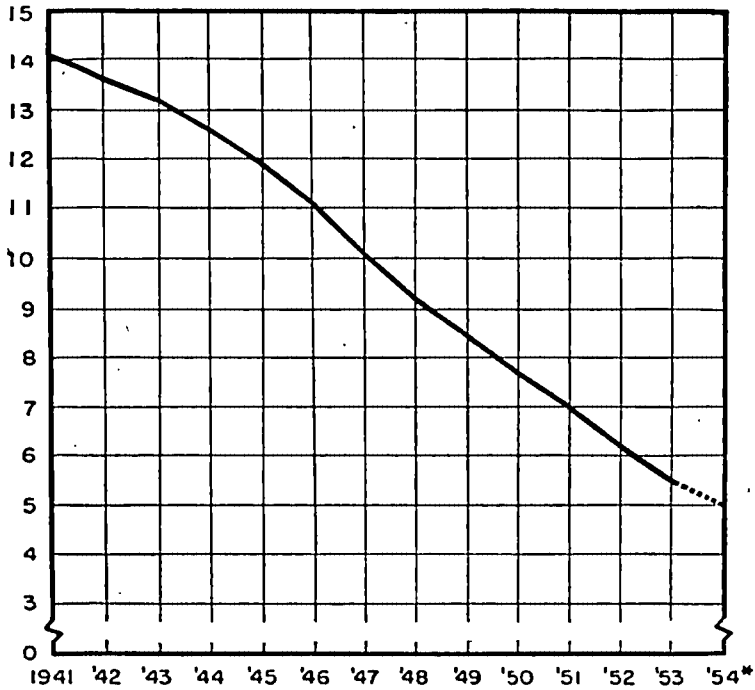
took place on a level different from that before. Some new levels were higher, some lower, than the earlier ones.

In such instances, it may not be possible for the statistician to represent the growth (or decline) as one trend in the series affected. It may then be appropriate to represent the growth factor on each level as a separate trend. Chart 14.4 illustrates such a situation, where a new level is reached after World War II for revenue passengers carried on domestic airlines in the United States.

Seasonal Variation

Certain movements which influence data through time recur at regular time intervals. These movements are called periodic

Horses and mules
in millions



*Preliminary

Chart 14.2. Horses and Mules on Farms in the United States.
1941-1954.

Source: Statistical Abstract of the United States, 1954.

Population
in thousands:

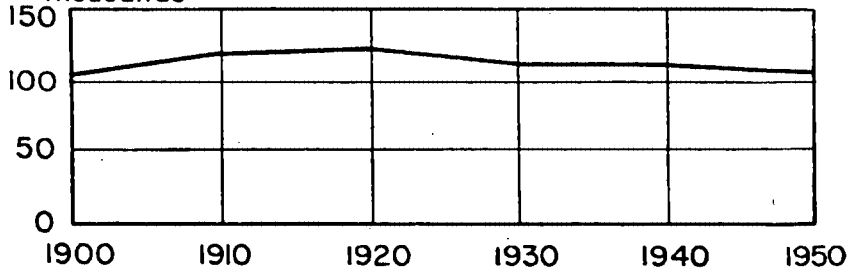


Chart 14.3. Population of Fall River, Massachusetts, 1900-1950.

Source: United States Bureau of the Census, 1950.

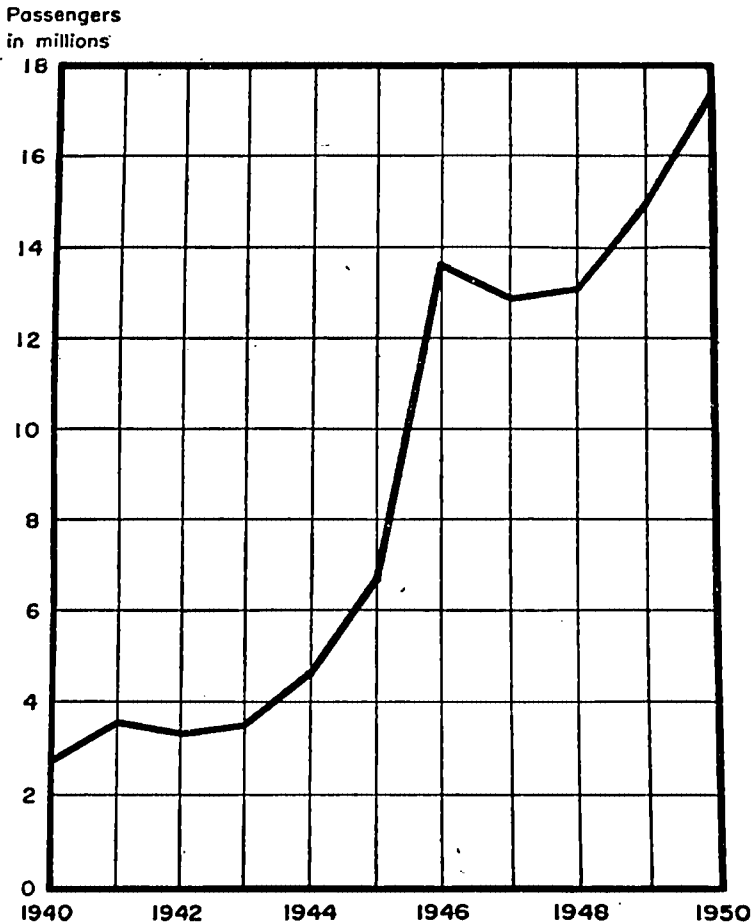


Chart 14.4. Revenue Passengers Carried on Domestic Air Lines in the United States, 1940-1950.

Source: Air Transport Association of America, 1951.

movements. Such movements may repeat themselves every day, for example the variation in the sales activity of a store with rush hours and slow periods; or every week, for example the variation in the business of a movie theater, with large receipts on week ends; or every month, as in the deposits at certain savings banks on the first of the month. The most

important type of these movements, however, is the one that recurs every year. This type of periodic movement is called a seasonal variation. Comparatively high retail sales appear before Christmas in every year.

Seasonal variations have two main causes: (1) climate in its widest sense, and (2) customs. Climate influences the timing of farmers' incomes. Sales of clothing have a seasonal movement due to climate. Customs determine the timing of such consumer expenditures as those at Christmas and Easter. Employment in certain industries, too, follows a seasonal pattern.

The establishment of the seasonal pattern which a series tends to follow year after year is the goal of seasonal analysis. This seasonal pattern for egg production in the United States is shown in Chart 14.5. This series, every year, follows approxi-

Seasonal Index
(in percent)

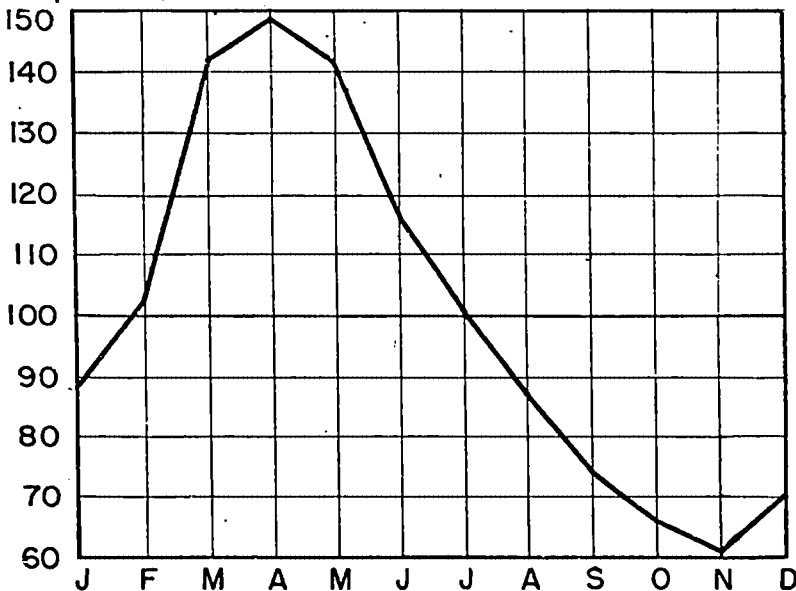


Chart 14.5. Seasonal Pattern in Egg Production in the United States, 1938-1947.

Source: *Survey of Current Business*; Supplements: 1942, 1947, 1949.

mately the same seasonal pattern; it shows a high in April and a low in November every year, regardless of total absolute production.

The seasonal pattern may remain the same or it may show changes in the long run. The introduction of the automobile sedan and the automobile heater years ago caused a *gradual* change in the seasonal pattern of automobile sales, since more and more cars could be sold in fall and winter because protection against inclement weather was afforded by the closed sedan with heating system. On the other hand, a change in the date of the annual exhibition of new automobile models caused an *abrupt* change in the seasonal pattern of automobile sales.

Cyclical Variation

Most economic series are influenced by the wavelike changes of prosperity and depression which have marked our economic system. In times of prosperity, production, sales, employment, and other economic activities are high; in times of depression the opposite is true. Thus these cycles of economic activity cause a wavelike movement of data through time. They show no regularity as to when they recur and how long they last, however; thus, they are to be distinguished from periodic movements.

What causes these cyclical movements? Unlike the causes of long-term and seasonal movements, we cannot easily establish the causes of cycles. Not only the causes of cyclical movements, but even the very concept of a cycle, its phases, and theories about its nature and duration, have been much debated in economic theory. Study of cycles makes us aware of the limitations attached to measuring this highly involved phenomenon. We cannot hope to do more than estimate a cycle, and predictions even of trained experts as to future movements of cycles have shown high degrees of inaccuracy. There are even doubts that there is anything that may be properly called a cyclical pattern.

Chart 14.6 illustrates the undulatory movement in the production of Portland cement from 1935 to 1940. It shows a cyclical high in 1936, and a cyclical low in 1938.

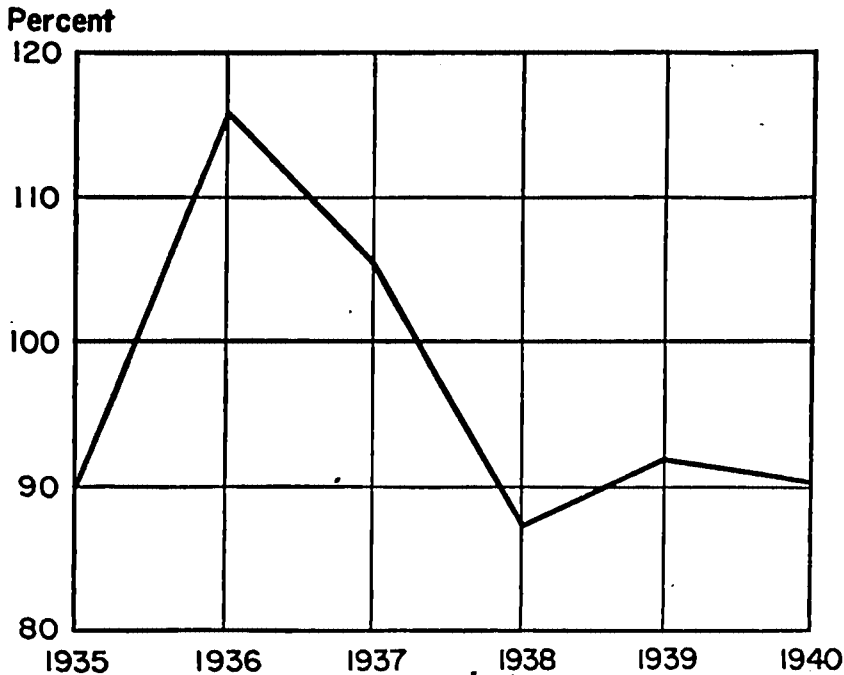


Chart 14.6. Cyclical Movements of Portland Cement Production in the United States, 1935–1940.

Source: *Survey of Current Business* (various issues), United States Department of Commerce.

Irregular Variation

Up to this point, we have discussed in broad terms three elements of the composite force which shapes a series through time: *trend* which depicts the inherent tendency to grow or to decline over a period of time; *seasonal variation* whereby the data conform to a seasonal pattern with highs and lows at different times of the year; and *cyclical variation* which represents the influence of “good” and “bad” times. At any point in time these three elements of the composite force are at work. Steel production as of August 4, 1955, for example, was determined by the long-term growth factor in the steel industry, by the seasonal position of August in the steel industry, and by the cyclical phase at this point in time.

In addition to these three elements, every series is subjected to occasional influences, which may occur just once, or several times, but without any pattern or other regularity. The variations they produce are therefore called *irregular variations*. A strike will push down production figures; a fire in a department store will influence sales and even employment data. Wars, earthquakes, floods, and other unforeseen or unforeseeable events are typical causes of these variations. An irregular variation may last but a day, or may last many months.

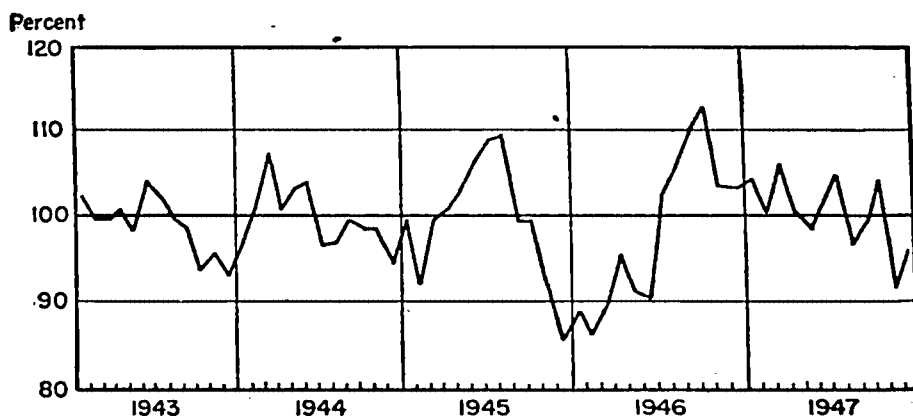


Chart 14.7. Irregular Variations in Factory Production of Creamery Butter in the United States, 1943-1947.

Source: Original Data Compiled by Bureau of Agricultural Economics, United States Department of Agriculture. Reported in: *Survey of Current Business*, Supplements 1947, 1949.

Chart 14.7 illustrates irregular variations in butter production from 1943 to 1947.

PREPARATION FOR ANALYSIS OF A TIME SERIES

Having defined what we are studying, our next step is to isolate and measure the four elements of the composite force which shapes a series in its motion through time.

A few words of caution are in order here. It has already been

mentioned that there is a subjective factor in any statistical work. This subjective factor is especially strong in time-series analysis. The statistician has to diagnose changes in terms of elements at work, and the way he analyzes the data depends on the diagnosis he has made. One statistician may look upon a given fluctuation as caused by the growth factor, and another statistician may see it as caused by a cycle. Furthermore, the statistician is attempting to measure complex economic phenomena; no precision may be expected in measurements of concepts whose exact definition is not agreed upon generally. Nevertheless, statistical analysis of time series is the alternative to mere guesswork, and is of value in economics and business *once its limitations are realized*. Although only estimates, our representations of the four elements of a time series have proved to be valuable working tools.

Editing Time-Series Data

The first step in time-series analysis is to insure comparability among the data. Is the total production for January comparable to the total production for February? Can we compare sales figures for 1944 and 1954?

Certain adjustments in the original data may be necessary. Here again sound judgment, good sense, and understanding of the subject matter must guide us. Adjustments may be needed for: (1) calendar variations; (2) price changes; (3) population changes; (4) miscellaneous changes. As is usually the case in statistics, we shall eliminate disturbing factors if we divide by them. These adjustments are discussed below.

1. Calendar variations. Taking a series of monthly sales data as an illustration, we see that they will frequently not be comparable because the number of days is not the same in each and every month. We can eliminate this difficulty by dividing the sales for each month by the number of days in this month. In some cases it will be preferable to express monthly data per calendar day: for example, the receipts of a movie theater; in other cases it will be preferable to express monthly data per

working day: for example, mean wages in an industry where the length of the work week varies during the year. Sometimes comparability of monthly data may be achieved by expressing them per week rather than per day.

2. *Price changes.* Sales value is quantity of units multiplied by price per unit. Since both quantities sold and prices at which they sell change from one time to another, no valid comparisons of sales can be made unless we eliminate the disturbing influence of changing prices. This elimination can be accomplished by dividing the sales figures for given time periods by the prices of the respective time periods.

This adjustment for price changes is called "deflating." * It is based on the following consideration:

$$\text{sales value} = \text{quantity} \times \text{price},$$

or

$$v = q \times p.$$

where v = sales value in dollars,

q = quantity of sales, and

p = price of each unit, in dollars.

The same equation, with a transposition, becomes

$$q = \frac{v}{p}.$$

The equation in this form enables us to find or estimate the number of units. Then we can compare the approximate quantities sold, since these quantities have been made comparable.

If sales figures refer to more than one item, a general measurement of "price" must be used. This will be found in a suitable price index (of which more will be said when we come to index numbers in Chapter 18).

* Adjustment for price changes is always called "deflating" although this term is not appropriate in time periods when prices are comparatively low. Then it may actually be "inflating" on a comparative basis.

3. *Population changes.* A comparison of the total meat consumption in the United States in 1915 and that in 1955 may not be useful since the number of consumers increased by many millions in those forty years. But if we divide each annual consumption figure by the respective population for the year we get meat consumption per person, or what is known as per capita consumption.

Sales data and many other types of data are often expressed on a per capita basis through making adjustment for population changes.

4. *Miscellaneous changes.* In practical work, we come upon changes which must be taken into account in making data comparable. The units in which the data are reported may have changed during the span of the time series, as for example, a change in reporting from long tons to short tons. The definition of a group of products being studied may change, and consequently the original data may be smaller in time periods when one product has been excluded by definition, or larger when a product has been added. For example, if production or sales of house dresses are being studied, we must be aware that the definition of house dress at one time included the Hoover apron, at another time excluded it. Definitions, classifications, and the types of product are subject to change, and these changes must be accounted for.

Many wrong conclusions can be avoided if the comparability of data is insured in a time series.

Graphic Presentation of Data

The next step toward analysis of a time series is to plot the comparable data. It is much easier to recognize developments in a series of data and to make decisions about the methods to be used in the analysis of a time series if the data are plotted on a graph.

It will depend on the particular study whether an arithmetic or semilogarithmic grid is appropriate. If there is any doubt, the data should be plotted on both types of grid.

We are now ready for measurement of the elements of the composite force. This measurement will be the subject of the next three chapters.

Summary

1. A study of the dynamics of data organized chronologically is the goal of time-series analysis.
2. There are four elements which cause variation in data over time: trend, seasonal, cyclical, and irregular variations.
3. Trend is long-term development which may be upward or downward.
4. Periodic movements which recur every year are called seasonal movements.
5. Cycles are wavelike movements reflecting prosperity or depression.
6. Fortuitous movements with no pattern or regularity are called irregular variations.
7. Before analyzing time series, certain preliminary steps have to be taken: namely, editing the data and plotting the data.

CHAPTER

15

Trend

Reasons for Trend Analysis

Given any long-term series, we wish to determine and present the direction which it takes—is it growing or declining? A graph of the *original data* gives us only a rough idea of the growth factor involved. It is possible, by computation, to measure this growth factor with some accuracy and thus arrive at a description of an underlying tendency. The direction known, we wish to establish the intensity of growth or decline over the long term—does this intensity remain the same all the time, or does it vary, being strong at one time, feeble at another? If the direction and intensity show constancy, then we may be able to represent the growth factor by a straight line. But changes in the direction or intensity cause bends in the line describing the growth factor.

It must be remembered that the measurement of trend, like the measurement of any element of the composite force, is on the level of *estimates*, not of precision. Moreover, for trend the data must be available for a long time span since in the short run the growth factor cannot be determined.

What are the reasons for measuring the tendency in a series to grow or decline over time? (1) To find out trend characteristics in and of themselves; (2) to enable us to eliminate trend in order to study other elements of what we have called the composite force.

1. In studying trend *in and of itself*, we ascertain the growth factor. For instance, we can compare the growth in the chemical industry with the growth in the economy as a whole, or with the growth in other industries; or we can compare the growth in one firm of the chemical industry with the growth in the industry as a whole. Thus, an investor may get a general idea of which company in the industry has shown greatest growth, and may accordingly invest his funds in that company as against others in the industry. In fact, in investment circles, certain industries are known as "growth industries" and certain companies as "growth companies." Moreover, we can compare through trend characteristics the growth of the chemical industry in the United States with that of other countries. The comparison of two trend lines is basically a comparison of their direction, and of their slope (that is, the amount of their increase or decrease over a unit of time).

Furthermore, assuming continuation of past trend, measurement of its characteristics may give us an indication of what is ahead for a given series. This prediction of future trend is called forecasting. Technically, this process of extending the trend into the future is known as extrapolation.

2. There are two purposes in *eliminating* trend. One is to get at the other elements of the composite force which influence data through time. To do this we must take trend out of the original data. The other purpose is to use data in the hypothetical form they would assume if trend were absent. This "adjustment for trend" is necessary in comparing or combining series with different growth factors.

The elimination of trend leaves us with seasonal, cyclical, and irregular factors. We can then, in two or more series, compare or use the impact of these three relatively short-term elements divorced from the long-term factor.

THE MEASUREMENT OF TREND

Trend can be determined: (1) by inspection or estimate; (2) by computation. Method 1 includes the freehand method

and the selected-points method. Method 2 includes the semi-average method, the least-squares method, and the moving-average method. The semi-average method partakes of both estimate and computation.

DETERMINING TREND BY INSPECTION OR ESTIMATE

The Freehand Method

Having made a graphic presentation of the original data in a series (a step preliminary to all time-series analysis, as we have seen), we may fit a trend line by inspection. We draw a line that, in our opinion, adequately describes on the graph the growth factor involved. This method obviously is highly

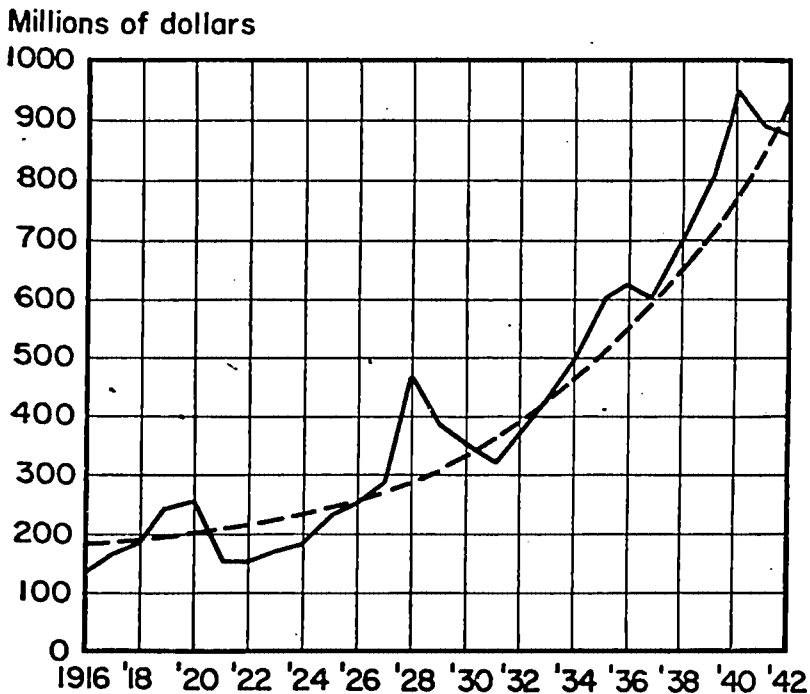


Chart 15.1. Trend Line Fitted by Inspection to Net Sales of Sears, Roebuck and Co., 1916-1942.

Source: *Moody's Manual of Investments*.

subjective, since the trend line depicts what the individual statistician sees as the trend. This freehand method should therefore be used only by experienced statisticians with a thorough understanding of the economic background of the particular series. Only after long experience in trend fitting should a statistician attempt to fit a trend line by inspection.

A trend line fitted by inspection for net sales of Sears, Roebuck and Company for the period 1916-1942 is shown in Chart 15.1.

The Selected-Points Method

We may select points, deemed characteristic, on or near the curve of the original data, and then connect the points. Thus we obtain a trend line that runs through a number of points considered typical of the growth factor observed in the series. If we think a straight line best describes the trend, then we need just two such characteristic points to plot it.

The selected-points method is really a refinement of the freehand method, though it provides marks to guide us. But like the freehand method it is highly subjective in that determination of typical points is left to the statistician.

DETERMINING TREND BY COMPUTATION: THE SEMI-AVERAGE METHOD

We may establish selected points by objective means; that is, we may find typical selected points by computation.

In the discussion of frequency distributions, it was pointed out that the arithmetic mean is a typical value and is representative of a series. If we break the time series we are studying into two equal parts, and represent each half by its mean, then a straight line passing through the two averages may be considered a rough description of the growth factor. This type of trend line is called a *semi-average** trend line. It is illustrated in Table 15.1 and Chart 15.2. In Table 15.1 we find that we have an odd number of years, namely, eleven. In order to

* "Semi average" means "average of semis" or "average of halves," that is, the average of each half of the series.

TABLE 15.1. COMPUTATION OF SEMI-AVERAGE TREND FOR ASSETS OF UNITED STATES LIFE-INSURANCE COMPANIES, 1943-1953.

Year	Assets, Billions of Dollars	Semi Average
1943	38	44.8
1944	41	
1945	45	
1946	48	
1947	52	
1948	56	68.8
1949	60	
1950	64	
1951	68	
1952	73	
1953	79	

Source: *Life Insurance Fact Book*, 1954, p. 58 (published by the Institute of Life Insurance, New York City).

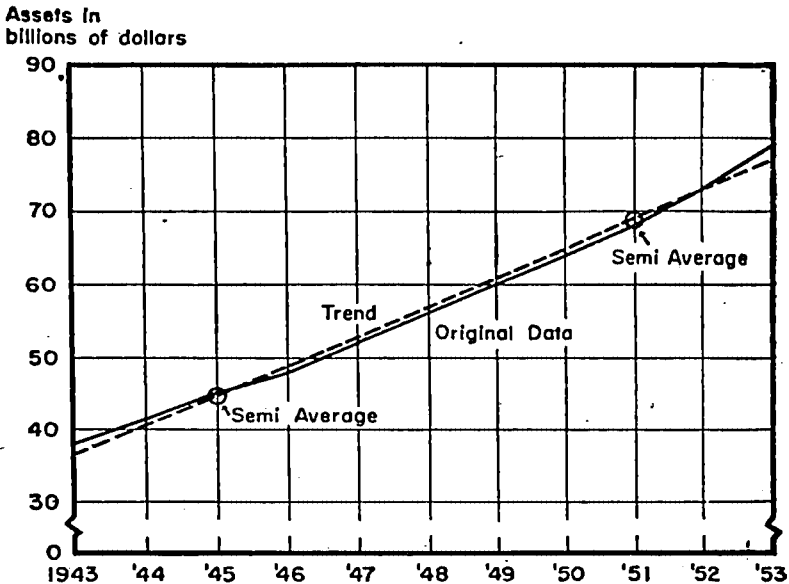


Chart 15.2. Semi-Average Trend for Assets of United States Life Insurance Companies, 1943-1953.

Source: Table 15.1.

divide this series into equal halves, we may either split the value in the middle year, here 1948, or add this value to each half of the series, or omit it. We have done the last. The two semi averages are plotted in the middle of their respective time spans; thus, in our illustration, 44.8 is plotted at 1945 and 68.8 at 1951.

The method of semi averages is not subjective, as are the inspection methods; but it suffers from the fact that the mean is greatly affected by extremes, and if there are extremes in either half or both halves of the series, then the trend line is not a true picture of the growth factor.

In many situations, simply plotting the semi averages and drawing a straight line will suffice to describe the growth factor. But if results obtained from this graph are not precise enough to establish trend values, to eliminate trend, or to forecast future trend, then we must find a way to establish trend in mathematical terms. We must therefore look for a general expression of a trend line which will permit us to find trend values for each time period in the data, and also for future time periods. How do we find such a general expression of a trend line?

DETERMINING TREND BY COMPUTATION: THE LEAST-SQUARES METHOD

Introductory Illustration

We first simplify the horizontal or X scale by substituting for the consecutive years the figures 0, 1, 2, and so on. We may take any year as zero, or what is called *the origin*, but here it is most convenient to call the first year the origin. Thus, for the life-insurance data in Chart 15.2, we call 1943 the origin and number it 0. At this origin of 1943, the trend line has a certain Y value; in our example this value is approximately 36.8. This Y value at the X origin is called the Y *intercept* and is designated by the letter a .

Every straight-line trend shows a steady change; the increment (or decrement) from one year to the next is always the same, and is called the slope. The symbol b is used to designate the

slope; b indicates how much the value of Y increases for each increase of one unit (in our case, one year) on the X -axis. In our illustration in Chart 15.2, this annual increment can be seen to be 4.0. The slope may be upward hence positive (as here), or downward, in which case it is negative.

Thus a , the Y intercept, tells us how high above the X -axis the trend line is at the origin; while b , the slope, shows us how much the trend line goes up or down as we move from one X unit to the next X unit. In Illustration 15.1, we see a and b schematically.

a and b are called *constants* because they remain unchanged within the series. But, of course, X and Y are not constants.

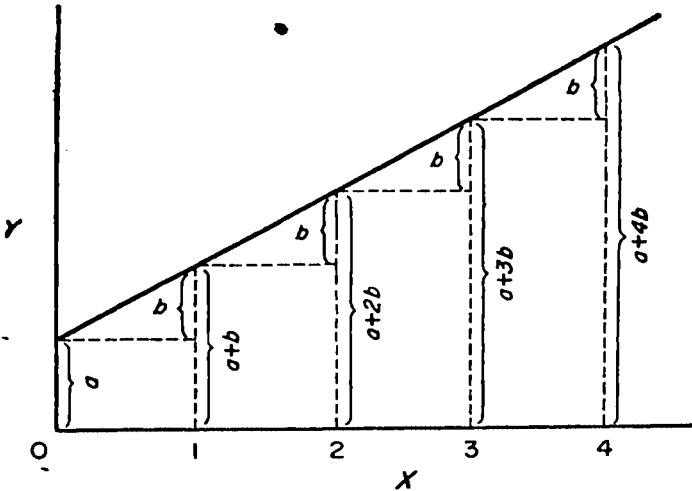


Illustration 15.1. Schematic Diagram of the Straight-Line Formula, $Y_c = a + bX$.

On the X -axis are the units of time, and this X variable is called the *independent* variable. That is, the Y variable (such as sales, production, employment, or prices) depends upon time. Therefore, the Y variable is called the *dependent* variable.

The original data are the Y values. If we plot the original Y values, we get, of course, a graph of the original data. But each Y value contains more than trend, and consequently for

each time period there is a difference between the Y value and the trend value for that time period. The Y values are given as original data, whereas the trend values are computed from them. Therefore the trend values are called computed values and symbolized as Y_c values.*

We may find these Y_c values through the basic formula for a straight line:

$$Y_c = a + bX$$

or, in our case,

$$Y_c = 36.8 + 4.0X.$$

The equation $Y_c = a + bX$ is the basis for measuring straight-line trend.

In our illustration we were able to take the value of a and the value of b from the graph. But to directly obtain trend values of greater accuracy, we must consider a and b as unknowns and we must compute them. We are thus led to the least-squares method.

Least-Squares Long Method

In a sales conference, though attendance was not taken, the secretary had to report the number present to the home office. He asked three salesmen who were there to venture a guess as to how many were in attendance. He got three different estimates—65, 70, 75. On the assumption that 75 and 65 were the extremes of possible actual attendance, the secretary asked himself: Which is the most probable figure? He then *minimized error* by taking the arithmetic mean, since only in the case of 70 is 5 the largest error that may occur in this instance. For any other figure, the error may be larger than 5. We already know from our study of the arithmetic mean that the deviations from it will balance, and that the sum of the squared deviations is a *minimum*. Therefore, the arithmetic mean is the best guess, or what we may call the “best fit” for a series of figures.

* Thus, what is termed O for each time period in the equation $O = T \times S \times C \times I$ is here called Y , and T for each time period is here called Y_c . Accordingly, $Y = Y_c \times S \times C \times I$.

In trend analysis, we face an analogous situation. We wish to represent the growth factor in the original data by a "line of best fit." In arriving at this line, our aim is likewise to minimize error; "error" here means deviation of each Y value from its corresponding trend value (see Illustration 15.2). The line we

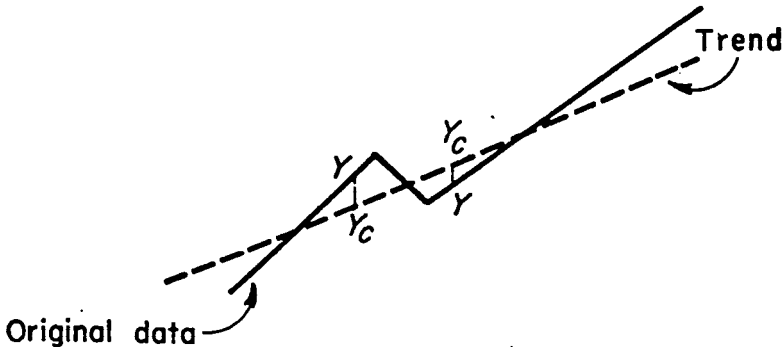


Illustration 15.2. Schematic Diagram Showing Deviations of Original Values (Y values) from Trend Values (Y_c values).

wish to find, must have the same characteristic as the arithmetic mean of the attendance at the sales conference—namely that the sum of all the squared deviations of the Y values from the Y_c values must be a minimum. A line that has this characteristic, namely $\Sigma(Y - Y_c)^2$ is least, is known as the *least-squares* line. Such a line will be a *dynamic average* of the movement through time of the original data, that is, a path that minimizes squared deviations and thus is an average path through the original data. Accordingly, this line must have another characteristic of an arithmetic mean, namely that the algebraic sum of the deviations of the items from the mean is zero, or in the case of this line, $\Sigma(Y - Y_c) = 0$.^{*} This average path through the original data is a *line of best fit*. How do we obtain this line? To fit this trend line to the original values (if a straight line best describes the growth factor), we have to know how far above the X -axis the line starts and how it ascends or descends from this starting point. In other words, we want to establish values for the Y

^{*} See pages 137 and 141.

intercept and for the slope. The first value is a , and the second is b ; in our basic equation $Y_c = a + bX$.

Since we have two unknowns (a and b), we must have two equations. We start from what are called "observation equations." Let us use for example a series of hypothetical employment data, which refer to a number of consecutive years. Here, as we did before, we can simplify X or time values by calling them 0, 1, 2, 3, 4, 5. We may designate as origin or 0 any year we wish, but for convenience we here take the first year in the series. Every year is one observation, and we may now set up our observation equations, following the pattern $Y = a + bX$, by substituting the given X and Y values for each observation.

<i>Year</i>	<i>Employment in Millions,</i>		<i>Observation Equation</i>
	<i>Y</i>	<i>X</i>	$(Y = a + bX)$
1950	2	0	$2 = a$
1951	3	1	$3 = a + 1b$
1952	4	2	$4 = a + 2b$
1953	6	3	$6 = a + 3b$
1954	5	4	$5 = a + 4b$
1955	10	5	$10 = a + 5b$

If we now take any two of these six observation equations, we can solve for a and b , but we get different values for a and b depending on which two equations we take. But what we are looking for are unvarying values for a and b . It is necessary, therefore, in order to obtain such values for a and b , to find what are called "normal equations." These normal equations come from summarizations of the observation equations after they have been manipulated. How must the observation equations be treated before summarization? We first multiply each of the observation equations by its coefficient of a and total them, and then multiply each of them by its coefficient of b and total them. These two totals are the "normal equations" we have been looking for.

In the equation $Y = a + bX$, the coefficient of a is 1, and

multiplication of the equations by it leaves them unchanged. Totaling them in our hypothetical employment data, we obtain

$$30 = 6a + 15b,$$

which is our first normal equation.

Multiplication of each original observation equation by the coefficient of its b , which is the value of X (0 or 1 or 2 or 3 or 4 or 5), gives us the following set of equations:

X	Observation Equation	Observation Equation Times X
0	$2 = a$	0
1	$3 = a + 1b$	$3 = 1a + 1b$
2	$4 = a + 2b$	$8 = 2a + 4b$
3	$6 = a + 3b$	$18 = 3a + 9b$
4	$5 = a + 4b$	$20 = 4a + 16b$
5	$10 = a + 5b$	$50 = 5a + 25b$
	Total	$99 = 15a + 55b$

The total shown is our second normal equation,

$$99 = 15a + 55b.$$

Simultaneous solution of these two normal equations gives us the values

$$a = 1.58 \quad \text{and} \quad b = 1.37.$$

Substituting these a and b values in the trend-equation $Y_e = a + bX$, we obtain for this hypothetical series

$$Y_e = 1.58 + 1.37X.$$

The results of our experiment with these hypothetical employment figures in finding a and b can be put in general terms. By expressing the normal equations in symbols, we get standard formulas for use in all practical work.

In our first normal equation, 30 is the summation of the Y values, or ΣY ; 6 is the number of observations, or N ; 15 is the

sum of the X values, or ΣX . Thus, $30 = 6a + 15b$ becomes, in general terms,

$$\Sigma Y = Na + b \Sigma X.$$

In the second normal equation, 99 was obtained as the sum of the product of the X and Y values, or ΣXY ; 15 is again the sum of the X values, or ΣX ; and 55 was obtained as the sum of the X values times the X values, or ΣX^2 . Thus, $99 = 15a + 55b$ becomes, in general terms,

$$\Sigma XY = a \Sigma X + b \Sigma X^2.$$

Thus, the two normal equations are

$$\Sigma Y = Na + b \Sigma X, \quad (I)$$

$$\Sigma XY = a \Sigma X + b \Sigma X^2. \quad (II)$$

Now we are ready to fit any straight-line trend by the least-squares method. We use these normal equations *without any further regard to their derivation from observation equations*. Let us fit the straight-line trend to the data in Table 15.2.

The trend equation for ore-production data in the Delta Copper Mines for the period 1947-1955 is found to be

$$Y_c = 12.08 + 0.73X.$$

Anybody reading just this equation would be in doubt about three facts: (1) Which year was selected as the origin? (2) What is the unit of time represented by X ? Is it five years, one year, a month, a half year? (3) In what kind of units is Y being measured? Is it production in tons, sales in millions of dollars, prices in dollars, employment in thousands of workers?

Hence, we supplement any trend equation with statements concerning the year of origin and the kind of units represented by X and by Y . Thus, the full and complete trend equation for the Delta Copper Mines reads as follows:

$$Y_c = 12.08 + 0.73X;$$

origin, 1947;

X units, one year;

Y units, production in thousands of tons.

TABLE 15.2. COMPUTATION BY LEAST-SQUARES LONG METHOD OF STRAIGHT-LINE TREND FOR ORE PRODUCTION IN THE DELTA COPPER - MINES, BUTTE, MONTANA, 1947-1955.

Year	Production in thousands of tons, Y	X	XY	X^2
1947	11	0	0	0
1948	13	1	13	1
1949	15	2	30	4
1950	14	3	42	9
1951	15	4	60	16
1952	16	5	80	25
1953	16	6	96	36
1954	17	7	119	49
1955	18	8	144	64
	$\Sigma Y = 135$	$\Sigma X = 36$	$\Sigma XY = 584$	$\Sigma X^2 = 204$

$$\Sigma Y = Na + b \Sigma X, \quad (I)$$

$$\Sigma XY = a \Sigma X + b \Sigma X^2, \quad (II)$$

or

$$135 = 9a + 36b, \quad (I)$$

$$584 = 36a + 204b. \quad (II)$$

We eliminate a by subtraction and obtain b :

$$584 = 36a + 204b \quad (II)$$

$$540 = 36a + 144b \quad (I)$$

$$44 = 60b \quad (II - I)$$

$$b = \frac{44}{60} = .73$$

Then we obtain a by substitution:

$$135 = 9a + 36(.73), \quad (I)$$

$$9a = 108.72,$$

$$a = 12.08.$$

Hence

$$Y_0 = 12.08 + .73X.$$

The trend equation for these data is now fully determined and may be used by anybody to find trend values.

Thus the trend value for 1950 is obtained by substituting 3 (that is, the X value for 1950) for X in the basic trend equation as follows:

$$Y_c = 12.08 + 0.73(3) = 14,270 \text{ tons.}$$

Least-Squares Short Method

Up to now, in the least-squares method, we have taken as the origin the first year of each series studied. We could also have taken any other year as the origin, and for that year X would be 0. Considerable saving of both time and effort is possible if the origin is taken in the middle of the whole time span covered by the entire series. The origin would then be located at the mean of the X values. Thus, in the series of all years 1947 through 1955 the mean is 1951. With 1951 having an X value of 0, the consecutive years starting from the mean in 1951 have the following X values: -1 for 1950, -2 for 1949, and so on; $+1$ for 1952, $+2$ for 1953, and so on. The sum of these X values is 0; that is, $\Sigma X = 0$, as can be seen schematically in Illustration 15.3.

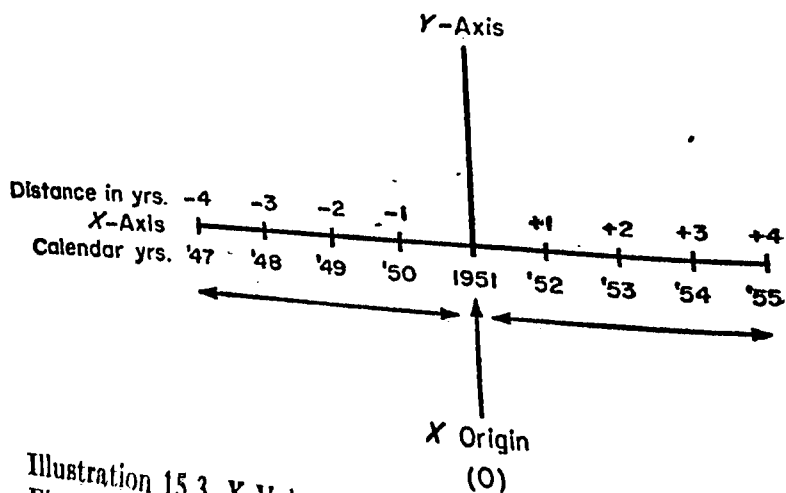


Illustration 15.3. X Values in Least-Squares Short Method for Finding Trend (Odd Number of Years).

If $\Sigma X = 0$, as here, then the two normal equations may be simplified. The first normal equation is

$$\Sigma Y = Na + b \Sigma X.$$

And since $\Sigma X = 0$, the equation becomes

$$\Sigma Y = Na \quad \text{or} \quad a = \frac{\Sigma Y}{N} = \bar{Y}.$$

The second normal equation is

$$\Sigma XY = a \Sigma X + b \Sigma X^2.$$

And since $\Sigma X = 0$, the equation becomes

$$\Sigma XY = b \Sigma X^2 \quad \text{or} \quad b = \frac{\Sigma XY}{\Sigma X^2}.$$

TABLE 15.3. COMPUTATION BY LEAST-SQUARES SHORT METHOD OF STRAIGHT-LINE TREND FOR ORE PRODUCTION IN THE DELTA COPPER MINES, BUTTE, MONTANA, 1947-1955.

Year	Production in thousands of tons, Y	X	XY	X ²
1947	11	-4	-44	16
1948	13	-3	-39	9
1949	15	-2	-30	4
1950	14	-1	-14	1
1951	15	0	0	0
1952	16	+1	+16	1
1953	16	+2	+32	4
1954	17	+3	+51	9
1955	18	+4	+72	16
	$\Sigma Y = 135$	$\Sigma X = 0$	$\Sigma XY = +44$	$\Sigma X^2 = 60$

$$a = \frac{\Sigma Y}{N} = \frac{135}{9} = 15; \quad b = \frac{\Sigma XY}{\Sigma X^2} = \frac{44}{60} = .73;$$

$$Y_o = 15 + .73X.$$

The new first normal equation, $a = \Sigma Y/N = \bar{Y}$, states that at the point of $\bar{X}$, the trend value is $\bar{Y}$. Or in other words, the trend line runs through the point of averages; that is, in our illustration, the trend value for 1951 ($\bar{X}$) is the mean of the production figures ($\bar{Y}$).

Since an infinite number of lines can run through this point of averages, we need to determine which of these lines best fits the data. This determination is provided by the second normal equation, $b = \Sigma XY/\Sigma X^2$, which establishes the slope.

Finding the trend equation by the short method in terms of least squares is illustrated in Table 15.3.

As can be seen from Table 15.3, the trend equation now is as follows:

$$Y_c = 15 + 0.73X;$$

origin, 1951;

X units, one year;

Y units, production in thousands of tons.

It is clear that at the 1951 origin the trend value is the same as the value of a since then X is 0. But a is still the Y intercept

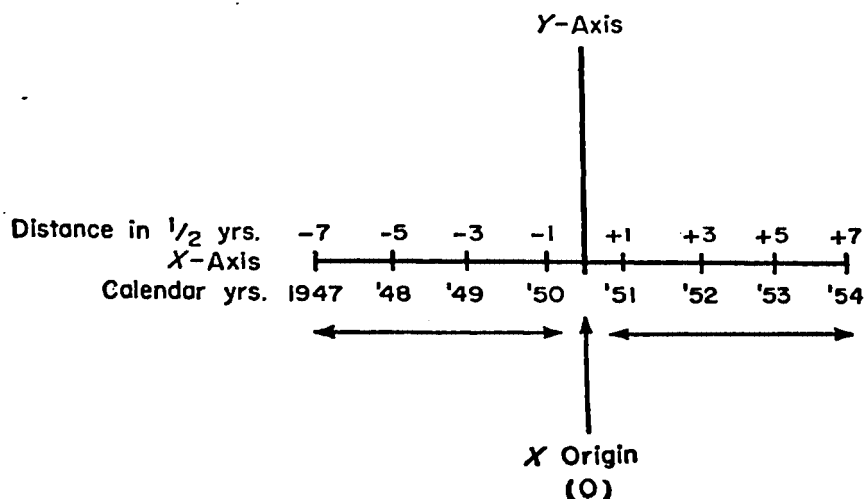


Illustration 15.4. X Values in Least-Squares Short Method for Finding Trend (Even Number of Years).

since the Y -axis now is located at 1951. This situation entails the use of the second quadrant for the $-X$ values and the first quadrant for the $+X$ values. (See page 43, Chapter 5.) b retains the same value as in the long method since the slope of a straight line does not vary.

Trend values remain the same for individual years regardless of whether the long or the short method is used. The trend value for 1947, which was 12.08 by the long method, is still 12.08 by the short method. But now the trend equation is stated with origin at 1951.

For odd numbers of time periods, the mean of the X values coincides with some time period. But if the time series has an even number of years, the mean falls between two years. Work

TABLE 15.4. COMPUTATION BY LEAST-SQUARES SHORT METHOD OF STRAIGHT-LINE TREND FOR NUMBER OF SHEEP ON FARMS IN THE UNITED STATES, 1942-1949.

Year	Number of sheep in millions, Y	X	XY	X^2
1942	56	-7	-392	49
1943	55	-5	-275	25
1944	51	-3	-153	9
1945	47	-1	-47	1
1946	42	+1	+ 42	1
1947	38	+3	+114	9
1948	35	+5	+175	25
1949	32	+7	+224	49
	$\Sigma Y = 356$	$\Sigma X = 0$	$\Sigma XY = -312$	$\Sigma X^2 = 168$

$$a = \frac{\Sigma Y}{N} = \frac{356}{8} = 44.5,$$

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{-312}{168} = -1.86,$$

$$Y_c = 44.5 - 1.86X.$$

Source: Bureau of Agricultural Economics, United States Department of Agriculture, 1950.

with such a mean would necessitate the use of decimals or fractions to measure the distance of any year from the origin. To simplify computations, we then talk in terms of half-year units instead of year units. Thus, for the years 1947–1954, the origin is *between* the years 1950 and 1951. The X value of 1950 (in half-year units) therefore is -1 , of 1949 is -3 , and so on; while the X value of 1951 is $+1$, of 1952 is $+3$, and so on. Thus, the sum of X values for even-numbered years also equals 0. (See Illustration 15.4, page 250.)

Finding the trend equation for time series with an even number of time periods is illustrated in Table 15.4 for the declining number of sheep on farms in the United States during the years 1942–1949.

Thus, the trend equation for these data is as follows:

$$Y_c = 44.5 - 1.86X;$$

origin, 1945–46;

X units, half year;

Y units, millions of sheep.

Number of sheep
in millions

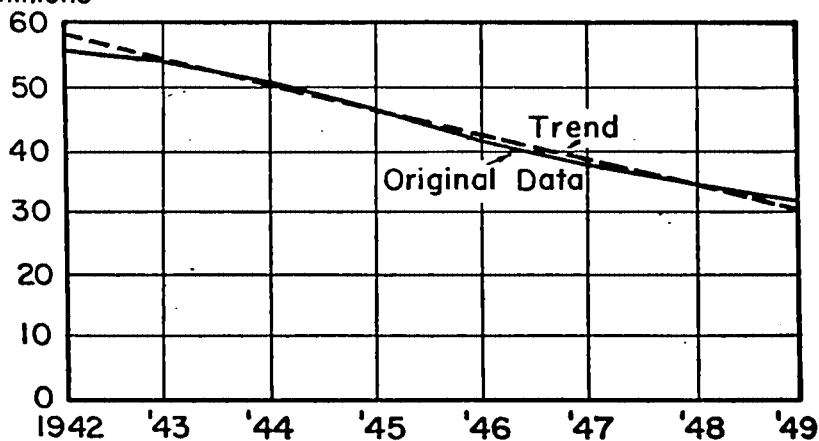


Chart 15.3. Straight-Line Trend Fitted to Number of Sheep on Farms in the United States, 1942–1949.

Source: Bureau of Agricultural Economics, United States Department of Agriculture, 1950.

The original data and the trend line are shown in Chart 15.3.

In the short method, for an even number of years, the slope involves half-yearly increases or decreases, and therefore b is half what the b value would be in annual units. But since the movement from one year to the next is $2X$ (in half-year units), the trend values arrived at from the trend equations are annual values. Thus the trend value for 1949 is $Y_c = 44.5 - 1.86(+7) = 31.5$.

USE OF THE TREND EQUATION

To what uses can the trend equation be put? It may be used to find any trend value *within* the range of time covered,

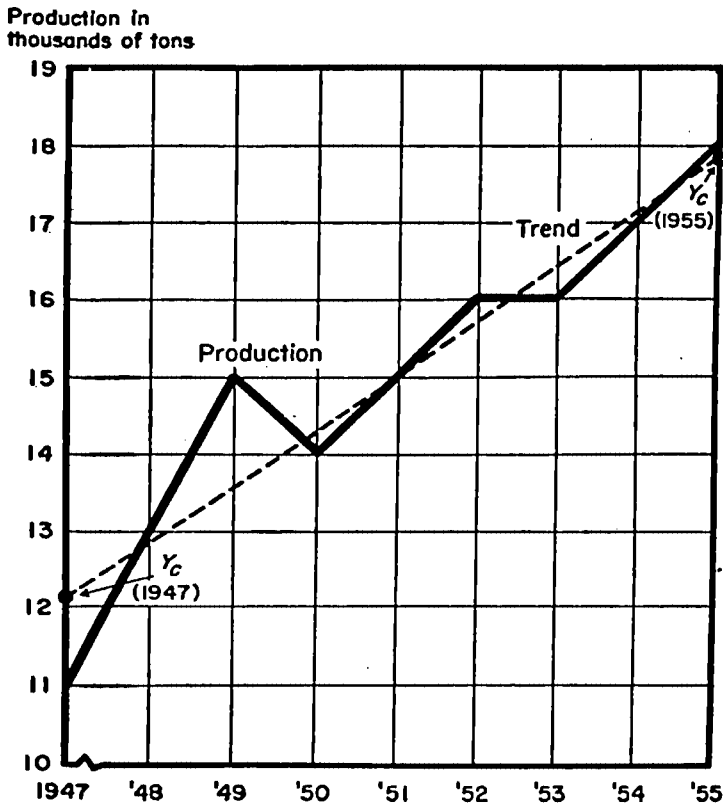


Chart 15.4. Straight-Line Trend Fitted to Ore Production in the Delta Copper Mines, Butte, Montana, 1947-1955.

or *beyond* * the range in either direction. To obtain a trend value for a given time period we ascertain that X value for this time period from our work sheet and substitute it for X in the trend equation, as we have already seen.

Further, we can plot the trend line by substituting for any two X values in the trend equation, by plotting the two trend values thus obtained at their respective time periods, and by passing a straight line through these two points. In drawing the straight line, it is preferable to take two time periods at some distance from each other on the graph, for example the first and the last. This wide spread lessens the possibility of inaccuracy in the drawing of the line. In Chart 15.4 is seen the trend line for the Delta Copper Mines, 1947-1955, arrived at by plotting and connecting the Y_c values for 1947 and 1955, and in Chart 15.3 is seen the trend line for the number of sheep on farms in the United States arrived at analogously.

Shifting the Origin

It is possible to shift the origin of a trend equation. The trend equation for the Delta Copper Mines is, as we know from Table 15.3:

$$Y_c = 15 + 0.73X;$$

origin, 1951;

X units, one year;

Y units, production in thousands of tons.

This origin of 1951 is assumed to be located at the middle of 1951 that is, July 1, 1951. Suppose we wish to state the trend equation with an origin at January 1, 1952. To do so requires, since X units are in whole years, that we add one-half of the annual increment, or $b(\frac{1}{2})$, to the trend value for July 1, 1951, as follows:

$$\begin{aligned} Y_c (\text{Jan. 1, 1952}) &= 15 + 0.73(\frac{1}{2}) \\ &= 15 + 0.365 \\ &= 15.37. \end{aligned}$$

* Finding such trend values is part of forecasting, and will be discussed in Chapter 17.

The trend equation now reads:

$$Y_c = 15.37 + 0.73X;$$

origin, January 1, 1952;

X units, one year;

Y units, production in thousands of tons.

In shifting from origin 1 to origin 2 as in Illustration 15.5,

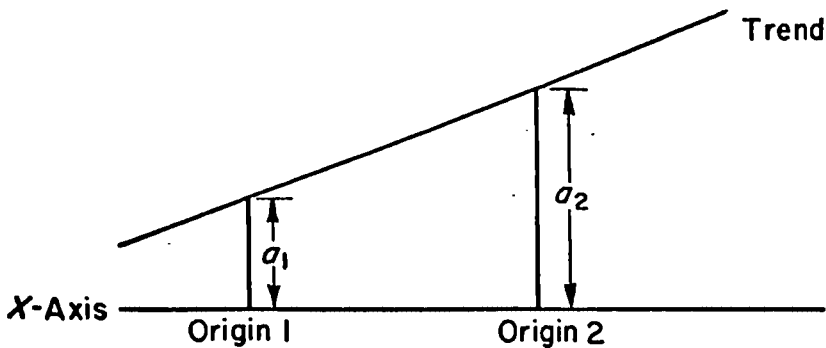


Illustration 15.5. Shifting the Origin of a Straight-Line Trend.

the a value changes but the b value remains the same, since the slope of a straight line does not undergo change. Shifting the origin therefore amounts to establishing a new a value. Since a is always the trend value at the point of origin, we have to find the trend value at origin 2 based on our trend equation with origin 1. In our case, the trend value at origin 2 is 15.37, b remains the same. We obtain the new trend equation by stating the new a value, 15.37, and repeating the b value.

Shifting the origin is of importance in converting a trend equation on an annual basis to a trend equation on a monthly basis. For refined analysis, especially of cycles, we may wish see cyclical turning points *during* a year; these are not noticeable in annual data. Of course, the trend line can be fitted to original monthly data by the same procedure that was illustrated for annual data; but this method may be laborious. It is usually easier to find the trend equation for corresponding

annual data and then convert the equation to the monthly level.*

THE METHOD OF MOVING AVERAGES

We can also fit a trend line by the method of moving averages.† But this method is not restricted to determining trend; it can also be used in connection with seasonal variations, cyclical variations, and irregular variations. In general, this method *tends to eliminate* any type of fluctuations by “ironing them out.” To iron out a type of fluctuations through moving averages means to remove their influence. Thus, the moving average is used as a tool in all parts of time-series analysis.

If we iron out cyclical fluctuations from annual data through a moving average, then the moving average for each year represents trend, since annual data do not show seasonality (for example, annual retail sales as a lump do not show Christmas or Easter business) and irregular variations are usually inconsequential in such data.

In general, the method of moving averages consists of taking the arithmetic mean of the Y values for a certain time span (number of years or of other time periods), and placing it at the center of the time span. Then, we repeat the procedure by dropping the first yearly figure of the Y values, adding on the figure directly following the last figure we had previously added; thus we move the time span and its center forward by one year; then we compute and place a new average. We continue until we exhaust the series. For example, we have annual production data, and we wish to find trend through a three-year moving average. The computation is shown in Table 15.5, and the resulting trend line in Chart 15.5.

* The procedure for converting an annual-trend equation to a monthly-trend equation is discussed in the Appendix to this chapter.

† While the computational methods heretofore discussed lead to standard types of curves (that is, curves capable of expression by simple mathematical formula), such is not the case with the method of moving averages. The moving-average trend curve is unique for each series, whereas least-squares curves fall into families.

TABLE 15.5. COMPUTATION OF TREND BY 3-YEAR MOVING AVERAGE FOR PRODUCTION IN DEEPWELL OIL CORPORATION, TULSA, OKLAHOMA, 1945-1954.

<i>Year</i>	<i>Production in Millions of Barrels</i>		<i>3-Year Moving Total</i>		<i>3-Year Moving Average</i>
1945	12	}	→ 39	→	13.00
1946	13				
1947	14				
1948	13		41		13.67
1949	14		42		14.00
1950	15		44		14.67
1951	15		46		15.33
1952	16		47		15.67
1953	16		47		15.67
1954	15				

What was the procedure followed in Table 15.5? First, we added the annual production figures for 1945, 1946, 1947, and secured a moving total of 39 million barrels. This was placed

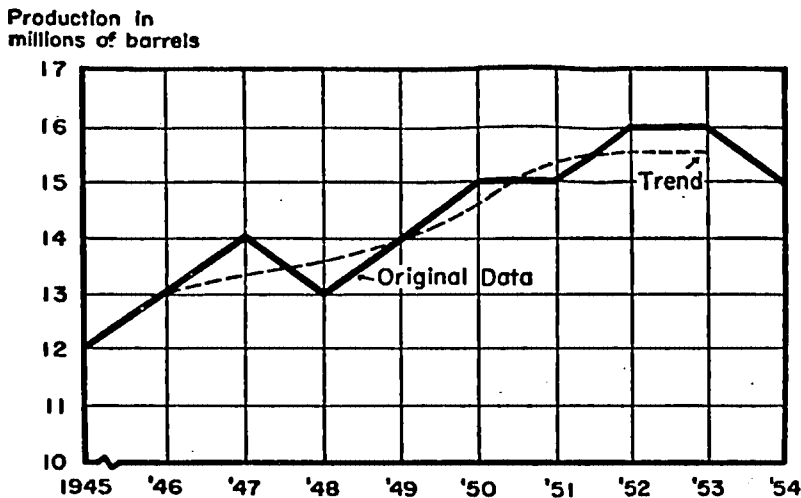


Chart 15.5. Trend Line by 3-Year Moving Average for Production in Deepwell Oil Corporation, Tulsa, Oklahoma, 1945-1954.

opposite the year 1946.* Then we dropped the value for 1945 and picked up the value for 1948. We thus got a moving total of 40 million barrels for 1946, 1947, 1948. This was placed opposite the year 1947. The process of dropping one year and picking up another was repeated throughout the series. (It will be noted that there cannot be a moving total opposite the first and the last year in such a case.) To secure moving averages we divided each three-year moving total by three. The series of three-year moving averages was taken to be an estimate of trend.

If the moving average is found for any odd number of time periods, then the method is analogous to the one shown above for the three-year moving average. For example, the first of a succession of five-year moving totals will be placed opposite the third year in a series, the next of the successive five-year moving totals at the fourth year in the series, and so on. Then each five-year moving total is divided by five.

If, however, the moving average is an even-period moving average—for instance, four-year or twelve-month—then the moving total and moving average which are placed at the center of the time span from which they are computed fall *between* two time periods. This placement is inconvenient † since the moving average so placed would not coincide with an original time period. We therefore synchronize moving averages and original data. This process is called *centering* and always consists of taking a two-period moving average of the moving averages. This process serves regardless of which even number we have in our period. A two-period moving total of the moving average is divided by two and gives us centered moving averages which now coincide with original time periods.

For example, if for the annual production data in Table 15.6, we take a four-year moving average, first we find a four-year

* A value representing a time span is considered to be located in the middle of the time span represented. Thus, an annual value is located at July 1 of the year it represents, and a three-year moving total is located in the center of the span represented.

† The inconvenience is mainly felt in plotting the trend values and in adjusting data for trend.

TABLE 15.6. COMPUTATION OF TREND LINE BY 4-YEAR MOVING AVERAGE FOR PRODUCTION IN THE DERRICK CORPORATION OF ST. LOUIS, Mo., 1944-1955.

<i>Year</i>	<i>Production (in Hundreds of Units)</i>	<i>4-Year Moving Total</i>	<i>4-Year Moving Average</i>	<i>2-Year Moving Total of 4-Year Moving Average</i>	<i>2-Year Moving Average of 4-Year Moving Average</i>
1944	5				
1945	6	25	6.25		
1946	7	26	6.50	12.75	6.38
1947	7	28	7.00	13.50	6.75
1948	6	30	7.50	14.50	7.25
1949	8	33	8.25	15.75	7.88
1950	9	36	9.00	17.25	8.63
1951	10	38	9.50	18.50	9.25
1952	9	40	10.00	19.50	9.75
1953	10	41	10.25	20.25	10.13
1954	11				
1955	11				

moving total for the years 1944, 1945, 1946, 1947. This we place *between* 1945 and 1946. We drop a year and pick up a year to get the other four-year moving totals. Then we find the four-year moving averages by dividing the moving totals by four. In order to center the four-year moving averages we compute a two-year moving average of them. The procedure can be seen from Table

15.6. The last columns in this table—"2-Year Moving Average of 4-Year Moving Average"—is the *centered 4-year moving average*. The resulting trend line is shown in Chart 15.6.

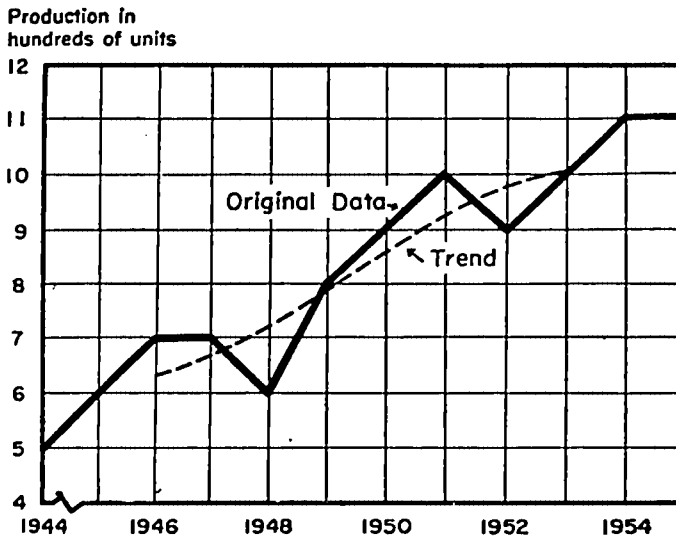


Chart 15.6. Trend Line by 4-Year Moving Average for Production in the Derrick Corporation of St. Louis, Missouri, 1944-1955.

How do we choose what time span to use in our moving average—three-year, six-month, four-year, twelve-month, or any other? This choice of span is determined by the length of time in the type of the fluctuation we are seeking to eliminate. For instance, cycles may last $2\frac{1}{2}$ years, 4 years, or some other period; irregular fluctuations may last 1 month or 4 months, or some other period. To iron out cyclical or irregular fluctuations we take a moving average based on the *average* duration of the fluctuation to be eliminated from the series. But seasonal patterns always have a duration of 12 months. The time span to be covered by the moving average for eliminating seasonal variations is thus determined.

Two factors prevent us from eliminating fluctuations *completely* in practical work. The first is that the amplitude of the

fluctuation to be eliminated varies; that is, its intensity varies. For example, depressions differ in severity and the amplitude of Christmas business varies from year to year. The second factor, which holds for cycles and irregular variations only, is that the duration of these fluctuations is not constant, and we have to take their average.

Limitations of the Moving-Average Method

We have already seen that the use of a moving average entails loss of information at both ends of a time series. And the longer the time span of the moving average the more information we lose. Thus, in a nine-year moving average we lose four years at each end, or a total of eight years. If we have a comparatively short time series, the losses may be so great as to make the use of a moving average inadvisable.

As a method of measuring trend, moving averages cannot be represented readily by a mathematical formula. Thus, this method is useful in eliminating trend but is not useful for comparison of trends and cannot be used to extend the trend line into the future.

ADJUSTMENT FOR TREND

The composite force at work in the original data was symbolized by the product $TSCI$. If we wish to get at the seasonal, cyclical, and irregular movements, we must first eliminate trend. To do so we must divide $TSCI$, the original data, by T for each time unit in the series.

In annual data, there is no S . In addition, I is usually negligible in annual data since irregular variations are usually short. Consequently, annual data are not symbolized by $TSCI$ but are usually approximated by TC .

To eliminate trend from annual data we divide the original data for each year, TC , by the corresponding trend value, T . The result gives us an estimate of C for each year (with some ir-

regular variations included). This procedure is called "adjustment for trend."

Adjustment for trend can be made in monthly or quarterly data as well. In that case, we divide the original data for each monthly or quarterly period, $TSCI$, by T for the period, and the result gives us SCI for each month or quarter.

This adjustment for trend may be done in order to: (1) study the other elements of a time-series (these other elements can be studied only after the long-term element has been eliminated); (2) use them to compare or combine series which are subject to different growth factors (for example, the *New York Times* Index of Business Activity uses steel and power data, each adjusted for trend).

CURVILINEAR TREND

We are here restricting ourselves to trends that can best be represented by a straight line. There are, however, time-series which are not properly represented by a straight line. Railroad-track miles went up from the Civil War through World War I, and then declined very slowly, from the peak. Such a development is best represented by a curvilinear trend.

Some curvilinear trends may straighten out on semilog paper. A semilog straight-line trend shows a constant *rate* of change. A different formula must be used to fit a straight line on semilog paper. An illustration of such a trend is given in Chart 15.7.

Most curvilinear trends do not plot as a straight line on either arithmetic or semilog paper. A simple, but rough method for fitting curvilinear trend is by breaking the series, not into two equal parts as in the semi-average method for straight-line trend, but into a larger number of equal parts depending on the number of important bends in the curve of the original data. We then take the average of each of these parts, plot these averages, and connect the points thus obtained. By least-squares, fitting curvilinear trend involves the introduction of one new unknown for each important bend, and this involves what are called

Sales in millions of dollars

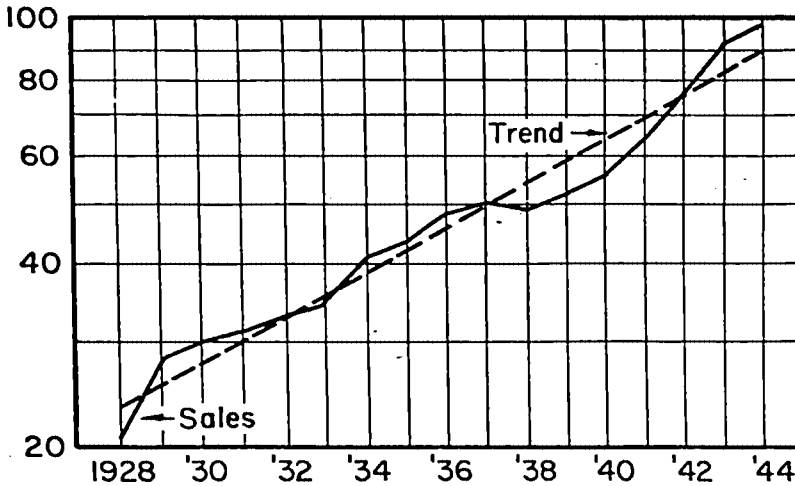


Chart 15.7. Trend Line Fitted to Sales of the J. J. Newberry Company, 1928–1944.

Source: *Moody's Manual of Investments*, 1945.

second-degree, third-degree, and further high-degree parabolas.* Thus, the least-squares method is applicable to fitting curvilinear trends as well as to fitting straight-line trends.

Summary

1. The study of the long-term growth factor in time series is made in order to find trend characteristics, or in order to permit the elimination of trend.

2. Trend may be determined by inspection or estimate (free-hand method or selected-points method); by the method of semi averages; by the least-squares method; by the method of moving averages.

* See the Appendix to this chapter for computation of nonlinear trend by least squares.

3. Inspection or estimate involves graphic approximation.
4. Semi averages are obtained by breaking the time series into two equal parts and finding the two averages. The least-squares method involves computing a and b , which are the constants in the straight-line trend equation.
5. The moving-average method is used to smooth out fluctuations of any type, as when we represent trend in annual data by ironing out cyclical fluctuations.
6. "Adjustment for trend" means eliminating trend from original data.
7. Where a straight trend line does not adequately represent the growth factor in a time series, we use a curvilinear trend line. The Appendix to this chapter discusses nonlinear trend fitted by least squares.

APPENDIX—SPECIAL PROBLEMS IN TREND ANALYSIS

CONVERSION OF ANNUAL TREND EQUATION TO MONTHLY TREND EQUATION

Fitting a trend line by least squares ($Y_c = a + bX$) to monthly data may be excessively time consuming. Thus, it is often more convenient to compute the trend equation from annual data and then convert this annual trend equation to a monthly trend equation. How is this done?

There are two different possible situations: (1) the Y units are annual totals, for example the total annual zinc production for the years 1955 to 1965; (2) the Y units are monthly averages, for example average monthly retail sales for the years 1955 to 1965. These monthly averages are the total annual sales for each year divided by 12.

Where Data Are Annual Totals

A trend equation operative on an annual level is to be reduced to a monthly level. The Y intercept or a value in the annual trend equation

is expressed in terms of annual Y values. To express the a value in terms of monthly Y values we must divide it by 12, thus transforming annual production to monthly production as regards the a value in the equation.

If we divide the slope b by 12, we reduce the annual change, let us say from 1955 to 1956, to a monthly change. But this division shows us only the change from some month in 1955 to the corresponding month in 1956, whereas what we are looking for is a change that expresses the difference between two consecutive months, for instance from January 1955 to February 1955. Therefore, b has to be divided by 12 once again. Obviously, it is much easier to divide b once by 144 instead of twice by 12.

Consequently, to convert an annual trend equation to a monthly trend equation when the annual data are expressed as annual totals, we divide a by 12 and b by 144. If the annual trend equation for the Pacific Expediting Corporation is

$$Y_c = 720 + 36X,$$

origin: 1950,

X units: one year,

Y units: total annual tonnage,

we can convert this equation to a monthly level, as follows:

$$Y_c = \frac{720}{12} + \frac{36}{144}X = 60 + .25X,$$

origin: July 1, 1950,

X units: one month,

Y units: monthly tonnage.

Where Data Are Given as Monthly Averages per Year

In this case, the Y values are, from the start, on a monthly level since they were obtained by dividing annual totals by 12. Therefore, the a value remains unchanged in the conversion process.

The b value in this case shows us the change on a monthly level, but from a month in one year to the corresponding month in the following

year. Here, therefore, it is necessary only to convert the b value to make it measure the change between consecutive months. Therefore, in this case we divide b by 12 just once.

Consequently, to convert an annual trend equation to a monthly trend equation when the annual data are expressed as monthly averages, we leave a unchanged and divide b by 12. If the annual trend equation for the Midcontinent Sales Corporation is

$$Y_c = 69 - 6X,$$

origin: 1953,

X units: one year,

Y units: average monthly sales,

we can convert this equation to a monthly level, as follows:

$$Y_c = 69 - \frac{6}{12}X = 69 - .5X,$$

origin: July 1, 1953,

X units: one month,

Y units: monthly sales.

Time Values in Half-Yearly Units

Up to this point, we have discussed the situation where the X units are one year in the annual trend equation. But if the X units are one-half year (as in series containing an even number of years; see pages 252-253), then the reduction is made not from an annual level to a monthly level but from a semiannual level to a monthly level.

Therefore, when X units are expressed in half years and Y units are annual totals, we divide a by 12 to bring it to a monthly level; and we divide b first by 6 and then by 12, or simply once by 72.

Consequently, to convert to a monthly trend equation an annual trend equation where the X units are half years and the Y units are annual totals, we do the following:

$$Y_c = \frac{a}{12} + \frac{b}{72}X.$$

If, on the other hand, X units are half years but Y units are monthly averages, then we leave a unchanged and we need divide b only by 6.

Consequently, to convert to a monthly trend equation an annual

trend equation where the X units are half years and the Y units are monthly averages, we do the following:

$$Y_c = a + \frac{b}{6}X.$$

Shifting the Origin

If the origin of an annual trend equation is, let us say, 1953, we take as the precise origin the center of this period, namely, July 1, 1953. If the origin is, let us say, 1951–1952, we take as the precise origin the center of this period, namely January 1, 1952.

A monthly trend equation must have its origin at the center of a month, that is, at the fifteenth of some month. Thus, a shifting of the origin becomes necessary whenever an annual trend equation is converted into a monthly trend equation.

If the origin of an annual trend equation is July 1, 1953, and we wish to state the monthly trend equation with origin at January 15, 1953, we substitute -5.5 for X in the monthly trend equation that has been obtained by conversion.

Shifting the origin has been discussed on pages 254–256.

NONLINEAR TREND BY LEAST SQUARES

When we plot original time series data on a graph preparatory to analysis, we may find that a straight line is not appropriate to the data but that a curved line may be, as is the case with the Sears, Roebuck and Company data on page 237. If there is one bend in the original data so that a curved line moving upward or downward will best represent trend, we have a so-called second-degree parabola. The trend equation in this case is

$$Y_c = a + bX + cX^2,$$

where a is the trend value at the time origin,

b is the slope at the origin, and

c establishes whether the curve is up or down and by how much.

The three normal equations needed to establish trend through a second-degree parabola are

$$\begin{aligned}\Sigma Y &= Na + b \Sigma X + c \Sigma X^2, \\ \Sigma XY &= a \Sigma X + b \Sigma X^2 + c \Sigma X^3, \\ \Sigma X^2 Y &= a \Sigma X^2 + b \Sigma X^3 + c \Sigma X^4.\end{aligned}$$

TABLE A15.1. COMPUTATION BY LEAST-SQUARES SHORT METHOD OF CURVILINEAR TREND FOR PRODUCTION
IN THE AMALGAMATED CORPORATION, CANTON, OHIO, 1944-1956.

Year	Production in tons, Y	X	XY	X^2	X^2Y	X^4	Y_c
1944	12	-6	- 72	36	432	1296	13.2
1945	20	-5	-100	25	500	625	12.4
1946	10	-4	- 40	16	160	256	12.2
1947	11	-3	- 33	9	99	81	12.7
1948	12	-2	- 24	4	48	16	13.8
1949	13	-1	- 13	1	13	1	15.5
1950	10	0	0	0	0	0	17.9
1951	21	+1	+ 21	1	21	1	20.9
1952	30	+2	+ 60	4	120	16	24.6
1953	35	+3	+105	9	315	81	28.9
1954	40	+4	+160	16	640	256	33.8
1955	37	+5	+185	25	925	625	39.4
1956	40	+6	+240	36	1440	1296	45.6
	$\Sigma Y = 291$	$\Sigma X = 0$	$\Sigma XY = +489$	$\Sigma X^2 = 182$	$\Sigma X^2Y = 4713$	$\Sigma X^4 = 4550$	

$$\Sigma Y = Na + c \Sigma X^2 \quad (1)$$

$$291 = 13a + 182c$$

$$\frac{291 - 182c}{13} = a$$

Substituting this value for a in equation (3), we find that $c = .32$. Therefore:

$$\begin{aligned} a &= \frac{291 - 182(.32)}{13} \\ &= \frac{291 - 58.24}{13} \\ &= 17.9 \end{aligned}$$

$$\Sigma XY = b \Sigma X^2 \quad (2)$$

$$489 = 182b$$

$$b = \frac{489}{182}$$

$$= 2.7$$

$$\Sigma X^2 Y = a \Sigma X^2 + c \Sigma X^4 \quad (3)$$

$$4713 = \left(\frac{291 - 182c}{13} \right) 182 + 4550c$$

$$4713 = (291 - 182c) 14 + 4550c$$

$$4713 = 4074 - 2548c + 4550c$$

$$639 = 2002c$$

$$.32 = c$$

Trend Equation

$$\begin{aligned} Y_c &= a + bX + cX^2 \\ &= 17.9 + 2.7X + .32X^2 \end{aligned}$$

In solving the above second-degree equations much time and labor can be saved by taking the time origin in the middle of the series so that $\Sigma X = 0$. But if $\Sigma X = 0$, then the sum of any odd power of X , such as ΣX^3 , is also zero. Therefore the three normal equations above become, when we take the origin in the middle of the series,

$$\begin{aligned}\Sigma Y &= Na + c \Sigma X^2, \\ \Sigma XY &= b \Sigma X^2, \\ \Sigma X^2 Y &= a \Sigma X^2 + c \Sigma X^4.\end{aligned}$$

For the illustrative data in Table A15.1, solving for a , b , and c gives us the basic trend equation for the data as follows:

$$Y_c = 17.9 + 2.7X + .32X^2$$

origin: 1950,

X units: one year,

Y units: production in tons.

The trend values for each year in this illustrative series are found by substituting in the trend equation the appropriate figures for X and X^2

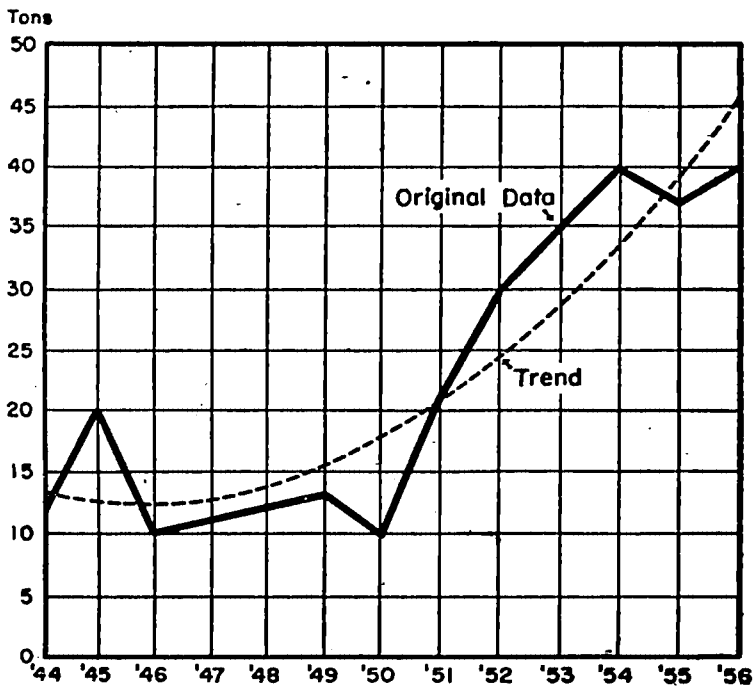


Chart A15.1

for each year. These trend values, in chronological order from 1944 to 1956, are: 13.2, 12.4, 12.2, 12.7, 13.8, 15.5, 17.9, 20.9, 24.6, 28.9, 33.8, 39.4, 45.6. In Chart A15.1 we see the original data and a second-degree parabola fitted to them.

The original data in certain cases may manifest more than one bend. In such a case we need a higher-degree parabola. Computing it involves adding an additional unknown for each additional bend. But a higher-degree parabola should be employed only when the statistician is confident that the bends reflect basic changes in the *growth* factor.

CHAPTER

16

Seasonal Variations

Reasons for Measuring Seasonal Variations

Why do we isolate the seasonal element? There are two major reasons: (1) to study seasonal variations in and of themselves; (2) to eliminate them.

1. By studying seasonal variations in themselves, we can get a clear idea about the relative position of each month in data relating to such matters as sales, production, employment, or the like. For example, in studying production data over time, analysis of the seasonal factor makes it possible to plan for the hiring of personnel for peak periods, to accumulate an inventory of raw materials, to ready equipment, and to allocate vacation time.

Seasonal variations in some industries and businesses are undesirable, and their measurement makes it possible to take action directed at leveling out these seasonal peaks and valleys in an enterprise (as when a manufacturer or retailer takes on a new line with seasonal fluctuations opposite to those in his current line). Labor unions are vitally concerned with seasonality in employment. Many industries are "highly seasonal"; and analysis of the seasonal pattern must precede decisions on how to overcome "seasonal unemployment."

In addition, whenever we forecast on a monthly or quarterly

basis, we must take account of the seasonal factor. A prediction of next October's sales figures is based not only on the trend factor but also on the seasonal position of October.

2. Why do we wish to eliminate the seasonal factor? In monthly or quarterly data, it is impossible to get at the cyclical or irregular factors until we isolate seasonality and eliminate it from the data. Moreover, in combining or comparing time series that have differing seasonal factors—for example, in comparing fur-coat and beach-wear sales in a department store, or in combining agricultural production and electric-power consumption for an index of business activity—we may want the data “deseasonalized,” that is, with the seasonal factor eliminated.

The Specific Seasonal and the Typical Seasonal

We find ups and downs due to seasonal factors in most economic time series. Of course, since seasonal variations recur every year, we cannot see the effects of seasonality if we lump together data by years or by longer time periods. Department-store sales for 1950 or 1956, for instance, do not show us the effects of Christmas business or Easter sales, as we know.

In quarterly, monthly, weekly, or daily data, seasonal variations are present. If we observe seasonal development during one year only, say 1952, we arrive at what is called a *specific seasonal*, namely that in 1952. If, however, we study seasonality for a number of years, we may come upon a pattern. Such a pattern is a generalized expression of seasonal variation for the series. The pattern thus is a *typical seasonal* obtained from a number of specific seasonals. The typical seasonal variation is, therefore, the *average* seasonal variation.

We obtain the typical seasonal for monthly data by averaging specific seasonals, that is, by averaging all Januarys within the span of time under investigation, then averaging all Februarys, and so on for each month of the year. If we have quarterly data, we must, of course, average all first quarters, all second

quarters, and so on. The averages, or typical January, typical February, and so on, constitute the seasonal pattern. This is the goal of seasonal analysis.

Certain adjustments have to be made, however, before we do this averaging. If we observe the sales data of a department store for one particular year, we may find that December sales are higher than the sales in January of the same year. The apparent reason is Christmas business. But is this the only explanation? The department store may show a tendency to grow over the years; thus, trend will have some part in lifting December sales over the sales of the preceding January. Furthermore, the year under scrutiny may be a year in the upward phase of a cycle. Thus cyclical movement tends to increase sales at the end of the year compared with those at its beginning. Finally, a one-time bonus to war veterans paid before Christmas may cause atypically high sales in this particular year; that is, irregular factors may lift December sales.

Thus, the fact that December sales for this particular year are higher than January sales is associated not only with the seasonal position of December, but also with trend, cyclical, and irregular factors. This realization helps us to outline our procedure: In order to isolate the seasonal pattern, we must first try to eliminate the disturbing influence of T , C , and I , as far as we can. Then we may average every January, February, and so on for the series.

Computation of Seasonal Variations

Several methods have been worked out to achieve the goal of isolating the seasonal factor. We shall present the one most widely used and generally considered satisfactory—the method of ratio to moving average.*

Let us illustrate the computation by the method of ratio to moving average for egg production in the United States for the years 1938 to 1947. Four steps are involved.

* The method of ratio to trend, and the link-relative method are also of some importance.

1. Our first step in determining the seasonal pattern is to *eliminate seasonality from the data*, despite the fact that its determination is our ultimate goal. We eliminate seasonality by ironing it out of the original data.

Since seasonal variations recur every year—that is, since the fluctuations have a time span of 12 months—a centered 12-month moving average tends to eliminate these fluctuations. This was discussed in Chapter 15. (In the case of quarterly data, a centered 4-quarter moving average must be used.) We cannot hope, however, to iron out seasonal fluctuations entirely, since their intensity varies over the years. Sales of skis may be very good in a snowy year and not good in a mild year. Thus we shall succeed only in eliminating the major part of the seasonal fluctuations. At the same time, a considerable portion of the short-range irregular variations will be smoothed out, too.

If we remain aware that we are speaking in approximations only, we may say that the centered 12-month moving average, which aims to eliminate seasonal and irregular fluctuations (*S* and *I*), represents the remaining elements of the original data, namely, trend and cycles. Thus, the centered 12-month moving average approximates *TC*.

The computation of a centered moving average has already been discussed on page 258. Thus, abstracted from Table 16.2 for egg production, we see that the original data and the centered 12-month moving average from July, 1938 to December 1939 are as shown overleaf.

It can readily be seen that the centered 12-month moving average here successfully irons out the great fluctuations caused by seasonal and irregular factors in the original data. The 12-month moving average shows the more gradual, longer range changes due to trend and cycles. Chart 16.1 shows the original data and the centered 12-month moving average for the entire series from 1938 to 1947.

2. The second step is to *take trend and cyclical fluctuations out of the original data*. We are then left with seasonal and irregular fluctuations. We must again emphasize that we are able only

		<i>Original Data</i>	<i>Centered 12-Month Moving Average</i>
		$T \times S \times C \times I$	$T \times C$
1938	J	2.45	
	F	3.02	
	M	4.53	
	A	4.90	
	M	4.57	
	J	3.73	
	J	3.24	3.12
	A	2.78	3.14
	S	2.32	3.14
	O	2.05	3.16
	N	1.75	3.16
	D	2.03	3.18
1939	J	2.63	3.18
	F	3.12	3.20
	M	4.62	3.20
	A	5.04	3.20
	M	4.76	3.22
	J	3.87	3.23
	J	3.31	3.23
	A	2.86	3.22
	S	2.40	3.22
	O	2.09	3.22
	N	1.88	3.23
	D	2.26	3.25

to approximate these four elements of a time series, and that we cannot represent them completely and precisely.

In terms of symbols, the second step therefore is as follows:

$$\frac{TSCI}{TC} = SI.$$

Our moving-average values represent TC . We divide them into the respective original egg-production data; for instance, for July 1938 we divide 3.12 into 3.24. The result, 1.038, is frequently called the "seasonal relative" and is expressed in percent. Thus, SI for July 1938 is 103.8%. The seasonal rela-

Production in
billions of eggs

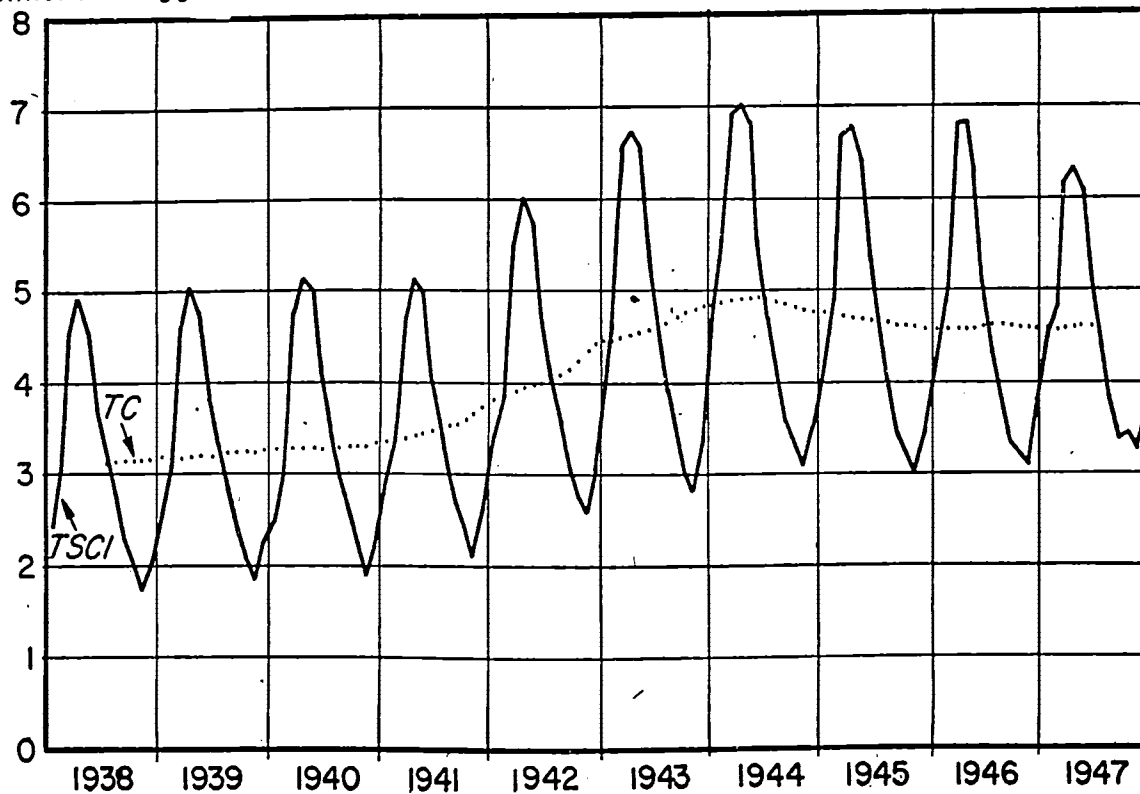


Chart 16.1. Egg Production in the United States, 1938-1947 (Original Data and 12-Month Moving Average).

Source: *Survey of Current Business*, Statistical Supplements 1942, 1947, 1949; also Table 16.2 herein.

		<i>Original Data</i> $T \times S \times C \times I$	<i>Centered 12-Month Moving Average</i> $T \times C$	<i>Seasonal Relatives</i> $S \times I$
1938	J	2.45		
	F	3.02		
	M	4.53		
	A	4.90		
	M	4.57		
	J	3.73		
	J	3.24	3.12	103.8
	A	2.78	3.14	88.5
	S	2.32	3.14	73.9
	O	2.05	3.16	64.9
	N	1.75	3.16	55.4
	D	2.03	3.18	63.8
1939	J	2.63	3.18	82.7
	F	3.12	3.20	97.5
	M	4.62	3.20	144.4
	A	5.04	3.20	157.5
	M	4.76	3.22	147.8
	J	3.87	3.23	119.8
	J	3.31	3.23	102.5
	A	2.86	3.22	88.8
	S	2.40	3.22	74.8
	O	2.09	3.22	64.9
	N	1.88	3.23	58.2
	D	2.26	3.25	69.5

tives from July 1938 to December 1939, abstracted from Table 16.2, are shown above.

Chart 16.2 shows the seasonal relatives for the entire series from 1938 to 1947. This curve shows the estimates of the seasonal and irregular factors combined.

We have now succeeded in eliminating from the original data to a considerable extent the disturbing influences of trend and cycles. It remains to rid the data of irregular variations. Then we shall be ready to average all Januarys, Februarys, and so on, and obtain the seasonal pattern.

3. The purpose of the third step is *to average, and—in the*

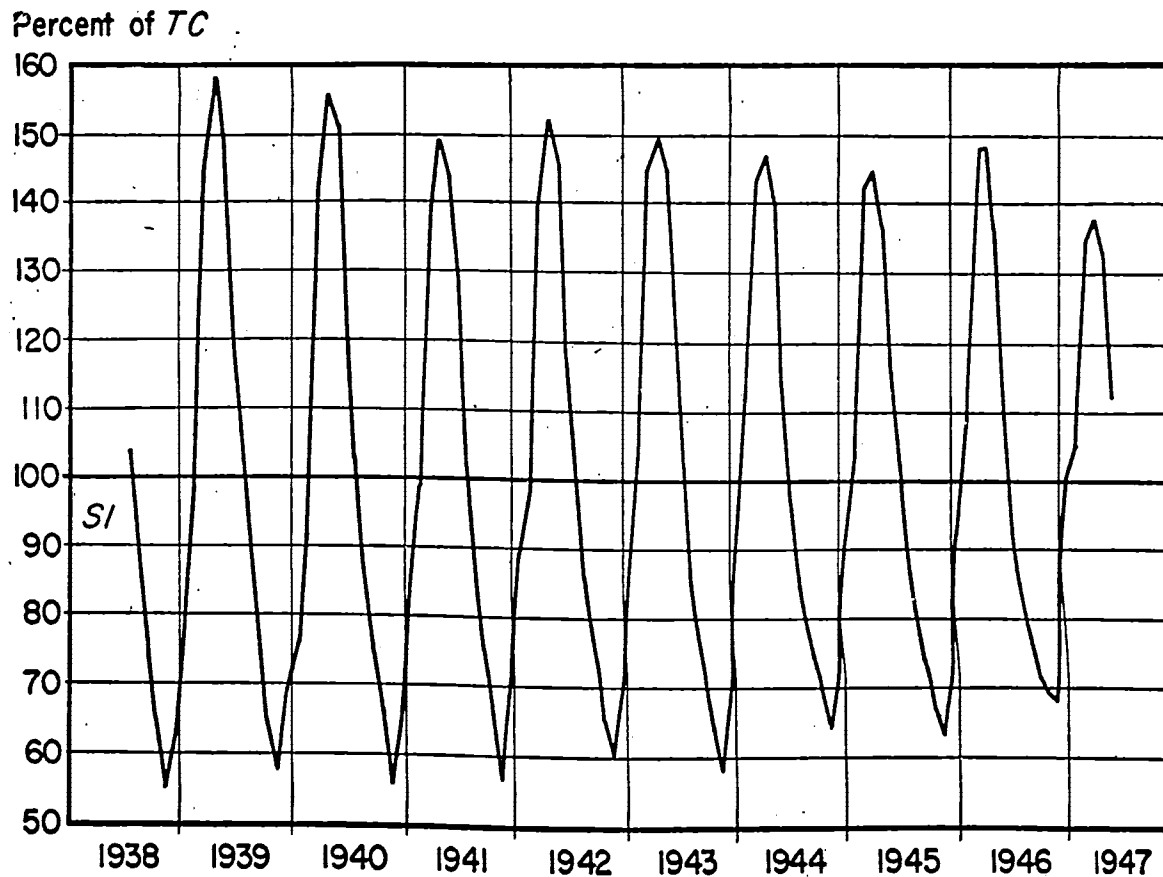


Chart 16.2. Seasonal Relatives for Egg Production in the United States, 1938-1947.

Source: Same as Chart 16.1.

process of averaging—to eliminate the irregular factor. We assume that the relatively high or extremely low values of seasonal relatives for any month are caused by irregular factors. The January low of 75.8 in 1940, for instance, may be due to an irregular factor. If we, therefore, exclude extreme values, we may hope to have eliminated the irregular element to a great extent.

This elimination of extremes may be achieved while we are averaging all Januarys, Februarys, and the like. We do this by using an appropriate type of average. We know the median is appropriate, since it is not affected by extremes. Thus, by using the median as an average, we can obtain the *typical seasonal relative* for each month, which will not be affected by irregular factors.

Sometimes a so-called modified mean is used as an average for each month. Here, extreme values are omitted before the arithmetic mean is taken. In an array of seasonal relatives for each month, a value or several values on one end or both ends may be dropped, and then the arithmetic mean of the remaining seasonal relatives is taken.

The third step in the computation of seasonal factors thus consists of arraying all January seasonal relatives, all February seasonal relatives, etc., then taking an average which eliminates extremes and hence eliminates the irregular factor. This process is shown in Table 16.1, where the median is used. The medians here are the values of the fifth item in the array of each month.

We have now obtained 12 typical seasonal relatives, one for each month. They are called the *crude seasonal index*. This index represents the seasonal element S in the purest form we can achieve. The typical seasonal relatives are expressed in percent. Thus, the March value of 142.1 means that this month is typically 42.1% above the trend-cycle value.

4. The fourth step is an adjustment *to eliminate certain small discrepancies*, such as those introduced in rounding. If we total the 12 medians from Table 16.1, we obtain 1203.3. But they should total 1200 or average 100; in consequence of rounding

and other operations, they come to slightly more. To reduce them to 1200, we multiply each month's typical seasonal relative by $1200/1203.3^*$ and thus reach a total of 1200. The adjustment in this case reduces each typical seasonal relative slightly downward. In some seasonal indexes, the total of the 12 medians (or modified means) may be less than 1200. In that case, the adjustment will slightly raise each typical seasonal relative.

This adjustment is made not only to achieve accuracy, but also because when we come to eliminate seasonality from the original data we do not wish to raise or lower the level of the data unduly. Thus, if a seasonal index aggregates more than 1200 (or averages more than 100), then the original data adjusted in terms of it will total less than the unadjusted original data. If it totals less than 1200, the opposite would be true.

The adjustment of the crude seasonal index for egg production in the United States results in the following:

January.	88.65
February	103.32
March	141.71
April	148.69
May	143.80
June	117.58
July	100.92
August	86.66
September	73.70
October	66.62
November	58.54
December	69.81
	1200.00

The adjustment of the crude seasonal index results in what is called the final seasonal index—the goal of our analysis.

This seasonal pattern has already been shown in Chart 14.5 (page 227).

The seasonal pattern in egg production clearly shows a seasonal high in April and a seasonal low in November. Moreover, the entire period from March through May is relatively

* Or .99726.

TABLE 16.1. ARRAYS OF PERCENTAGES OF CENTERED TWELVE-MONTH MOVING AVERAGE (SEASONAL RELATIVES) FOR EGG PRODUCTION IN THE UNITED STATES, JULY 1938 TO JUNE 1947, AND COMPUTATION OF SEASONAL INDEX.

Rank	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1	100.2	114.0	147.6	157.5	151.5	122.7	103.8	89.0	75.8	69.8	68.5	83.0
2	95.2	108.9	145.1	155.2	147.8	119.8	103.3	88.8	75.6	69.7	64.5	74.1
3	92.5	105.5	144.4	151.8	144.9	118.3	102.5	88.5	74.5	67.2	63.7	73.8
4	89.6	103.6	143.5	149.8	144.5	118.2	102.3	87.9	74.3	67.1	60.9	70.3
5 →	88.9	103.6	142.1	149.1	144.2	117.9	101.2	86.9	73.9	66.8	58.7	70.0
6	86.4	99.5	141.8	147.8	139.9	115.0	99.8	85.2	73.3	65.3	58.2	69.5
7	85.8	99.1	140.4	146.9	136.9	113.1	99.6	84.4	73.0	64.9	57.6	68.5
8	82.7	97.5	138.5	144.5	136.7	112.6	96.9	84.1	71.4	64.9	56.4	65.6
9	75.8	91.5	134.5	137.5	133.3	110.2	92.0	79.1	71.1	64.2	55.4	63.8
Median	88.9	103.6	142.1	149.1	144.2	117.9	101.2	86.9	73.9	66.8	58.7	70.0

Source: Derived from data in Table 16.2, pp. 283-285.

TABLE 16.2. COMPUTATION OF PERCENTAGES OF CENTERED 12-MONTH MOVING AVERAGE (SEASONAL RELATIVES) FOR EGG PRODUCTION IN THE UNITED STATES, 1938-1947 [PRODUCTION IN BILLIONS OF EGGS].

Year and month	Original data, TSCI	12-month moving total	12-month moving average (col. 3 ÷ 12)	2-month moving total of col. 4	Centered 12-month moving average (col. 5 ÷ 2) TC	Percent of centered 12-month moving average (col. 2 ÷ col. 6) SI
1	2	3	4	5	6	7
1938 J	2.45					
F	3.02					
M	4.53					
A	4.90					
M	4.57					
J	3.73					
J	3.24	37.37	3.11			
A	2.78	37.55	3.13	6.24	3.12	103.8
S	2.32	37.65	3.14	6.27	3.14	88.5
O	2.05	37.74	3.15	6.29	3.14	73.9
N	1.75	37.88	3.16	6.31	3.16	64.9
D	2.03	38.07	3.17	6.33	3.16	55.4
1939 J	2.63	38.21	3.18	6.35	3.18	63.8
F	3.12	38.28	3.19	6.37	3.18	82.7
M	4.62	38.36	3.20	6.39	3.20	97.5
A	5.04	38.44	3.20	6.40	3.20	144.4
M	4.76	38.48	3.21	6.41	3.20	157.5
J	3.87	38.61	3.22	6.43	3.22	147.8
J	3.31	38.84	3.24	6.46	3.23	119.8
A	2.86	38.68	3.22	6.46	3.23	102.5
S	2.40	38.56	3.21	6.43	3.22	88.8
O	2.09	38.59	3.22	6.43	3.22	74.5
N	1.88	38.67	3.22	6.44	3.22	64.9
D	2.26	38.91	3.24	6.46	3.23	58.2
1940 J	2.47	39.09	3.26	6.50	3.25	69.5
F	3.00	39.21	3.27	6.53	3.26	75.8
M	4.65	39.33	3.28	6.55	3.28	91.5
A	5.12	39.47	3.29	6.57	3.28	141.8
M	5.00	39.63	3.30	6.59	3.30	155.2
J	4.05	39.65	3.30	6.60	3.30	151.5
J	3.43	39.60	3.30	6.60	3.30	122.7
A	2.98	40.02	3.34	6.64	3.32	103.3
S	2.54	40.37	3.36	6.70	3.35	89.0
O	2.25	40.43	3.37	6.73	3.36	75.6
N	1.90	40.41	3.37	6.74	3.37	66.8
D	2.21	40.38	3.37	6.74	3.37	56.4
1941 J	2.89	40.42	3.37	6.74	3.37	65.6
F	3.35	40.57	3.38	6.75	3.38	85.5
M	4.71	40.71	3.39	6.77	3.38	99.1
		40.90	3.41	6.80	3.40	138.5

Table 16.2 (Continued)

Year and month	Original data, <i>TSCI</i>	12-month moving total	12-month moving average (col. 3 ÷ 12)	2-month moving total of col. 4	Centered 12-month moving average (col. 5 ÷ 2) <i>TC</i>	Percent of centered 12-month moving average (col. 2 ÷ col. 6) <i>SI</i>
1	2	3	4	5	6	7
1942	A	5.10	41.12	3.43	6.84	149.1
	M	4.97	41.38	3.45	6.88	144.5
	J	4.09	41.78	3.48	6.93	118.2
	J	3.58	42.32	3.53	7.01	102.3
	A	3.12	42.85	3.57	7.10	87.9
	S	2.73	43.67	3.64	7.21	75.8
	O	2.47	44.58	3.72	7.36	67.1
	N	2.16	45.39	3.78	7.50	57.6
	D	2.61	45.39	3.78	7.62	68.5
	J	3.43	46.05	3.84	7.72	88.9
	F	3.88	46.58	3.88	7.80	99.5
	M	5.53	47.03	3.92	7.87	140.4
1943	A	6.01	47.35	3.95	7.92	151.8
	M	5.78	47.66	3.97	7.98	144.9
	J	4.75	48.13	4.01	8.06	117.9
	J	4.11	48.60	4.05	8.13	101.2
	A	3.57	48.99	4.08	8.22	86.9
	S	3.05	49.73	4.14	8.37	73.0
	O	2.78	50.70	4.23	8.52	65.3
	N	2.63	51.43	4.29	8.64	60.9
	D	3.08	52.17	4.35	8.75	70.3
	J	3.82	52.79	4.40	8.84	86.4
	F	4.62	53.25	4.44	8.91	103.6
	M	6.50	53.58	4.47	8.96	145.1
1944	A	6.74	53.87	4.49	9.00	149.8
	M	6.52	54.12	4.51	9.03	144.2
	J	5.37	54.28	4.52	9.07	118.3
	J	4.57	54.54	4.55	9.16	99.8
	A	3.90	55.27	4.61	9.29	84.1
	S	3.34	56.12	4.68	9.39	71.1
	O	3.03	56.52	4.71	9.45	64.2
	N	2.79	56.89	4.74	9.50	58.7
	D	3.34	57.17	4.76	9.54	70.0
	J	4.55	57.32	4.78	9.57	95.2
	F	5.47	57.46	4.79	9.59	114.0
	M	6.90	57.63	4.80	9.62	143.5
1944	A	7.11	57.85	4.82	9.67	146.9
	M	6.80	58.14	4.85	9.67	139.9
	J	5.52	58.40	4.87	9.72	113.1
	J	4.71	58.53	4.88	9.75	96.9
	A	4.07	58.53	4.85	9.73	84.4
	S	3.56	57.58	4.80	9.65	74.3
	O	3.32	57.33	4.78	9.58	69.7
			56.97	4.75	9.53	

Table 16.2 (Continued)

Year and month		Original data, <i>TSCI</i>	12-month moving total	12-month moving average (col. 3 ÷ 12)	2-month moving total of col. 4	Centered 12-month moving average (col. 5 ÷ 2) <i>TC</i>	Percent of centered 12-month moving average (col. 2 ÷ col. 6) <i>SI</i>
1		2	3	4	5	6	7
1945	N	3.05			9.46	4.73	64.5
	D	3.47	56.55	4.71	9.41	4.70	73.8
	J	4.21	56.39	4.70	9.39	4.70	89.6
	F	4.86	56.32	4.69	9.38	4.69	103.6
	M	6.65	56.23	4.69	9.37	4.68	142.1
	A	6.75	56.10	4.68	9.34	4.67	144.5
	M	6.38	55.93	4.66	9.32	4.66	136.9
	J	5.36	55.86	4.66	9.31	4.66	115.0
	J	4.64	55.85	4.65	9.31	4.66	99.6
	A	3.98	55.93	4.66	9.34	4.67	85.2
	S	3.43	56.10	4.68	9.37	4.68	73.3
	O	3.15	56.24	4.69	9.38	4.69	67.2
1946	N	2.98	56.29	4.69	9.37	4.68	63.7
	D	3.46	56.20	4.68	9.34	4.67	74.1
	J	4.29	55.93	4.66	9.29	4.64	92.5
	F	5.03	55.57	4.63	9.24	4.62	108.9
	M	6.79	55.27	4.61	9.21	4.60	147.6
	A	6.80	55.14	4.60	9.20	4.60	147.8
	M	6.29	55.18	4.60	9.21	4.60	136.7
	J	5.09	55.31	4.61	9.25	4.62	110.2
	J	4.28	55.62	4.64	9.25	4.62	110.2
	A	3.68	55.89	4.66	9.30	4.65	92.0
	S	3.30	55.67	4.64	9.30	4.65	79.1
	O	3.19	55.04	4.59	9.23	4.62	71.4
1947	N	3.11	54.55	4.55	9.14	4.57	69.8
	D	3.77	54.39	4.53	9.08	4.54	68.5
	J	4.56	54.49	4.54	9.07	4.54	83.0
	F	4.81	54.73	4.56	9.10	4.55	100.2
	M	6.16	54.87	4.57	9.13	4.56	105.5
	A	6.31	54.93	4.58	9.15	4.58	134.5
	M	6.13	55.17	4.60	9.18	4.59	137.5
	J	5.19	55.32	4.61	9.21	4.60	133.3
	J	4.52	55.27	4.61	9.22	4.61	112.6
	A	3.82					
	S	3.36					
	O	3.43					
	N	3.26					
	D	3.72					

Source: *Survey of Current Business*, Statistical Supplements, 1942, 1947, 1949.

high, whereas the period from October through December is relatively low. Thus, the supply side of the egg market has been clearly marked out relative to the different months of the year. We may expect egg prices to be determined by these seasonal fluctuations of supply.

What are some practical consequences of such statistical analysis? An egg broker or wholesaler who puts aside eggs in cold storage should be prepared to begin accumulation in March, and continue through April and May. In addition, the seasonal pattern for eggs may be compared with the seasonal pattern for other foodstuffs. All planning on a monthly basis in egg production must be guided by the seasonal pattern which egg production tends to follow year after year.

ADJUSTMENT FOR SEASONALITY

To adjust data for seasonality means to deseasonalize the data. This adjustment is made by dividing the original data for each month by the corresponding seasonal-index value for that month; or in symbols,

$$\frac{TSCI}{S} = TCI.$$

The seasonal-index value for January is the same for every January in the entire series, and so on. Deseasonalized data for the years 1938 and 1939 are as follows:

		<i>TSCI</i>	<i>S</i>	<i>TCI</i>
1938	January	2.45	88.65%	2.76
	February	3.02	103.32	2.92
	March	4.53	141.71	3.20
	April	4.90	148.69	3.29
	May	4.57	143.80	3.18
	June	3.73	117.58	3.17
	July	3.24	100.92	3.21
	August	2.78	86.66	3.21
	September	2.32	73.70	3.15
	October	2.05	66.62	3.08
	November	1.75	58.54	2.99
	December	2.03	69.81	2.91

		<i>TSCI</i>	<i>S</i>	<i>TCI</i>
1939	January	2.63	88.65	2.97
	February	3.12	103.32	3.02
	March	4.62	141.71	3.26
	April	5.04	148.69	3.39
	May	4.76	143.80	3.31
	June	3.87	117.58	3.29
	July	3.31	100.92	3.28
	August	2.86	86.66	3.30
	September	2.40	73.70	3.26
	October	2.09	66.62	3.14
	November	1.88	58.54	3.21
	December	2.26	69.81	3.24

Summary

1. Seasonal variations are measured either to study them in themselves, or to eliminate them.

2. The method of ratio to moving average involves four steps in reaching the seasonal index. The first step is to iron out seasonality from the original data by a centered 12-month moving average which approximates *TC*.

3. The next step is to take out trend-cycle by dividing the centered 12-month moving average into the original data. The result gives us the seasonal relatives, or an estimate of *SI*.

4. The third step involves two different purposes: the elimination of the irregular factor, and the averaging of the seasonal relatives referring to the same month (or quarter). Thus we obtain the crude seasonal index.

5. The fourth step consists of adjusting, if necessary, the crude seasonal index, thus obtaining the final seasonal index.

6. To eliminate seasonality we divide the original data *TSCI* in each month (or quarter) by the corresponding value in the seasonal index *S*. We thus obtain *TCI*, or deseasonalized data.

CHAPTER

17

Cyclical and Irregular Variations; Forecasting

THE PROBLEM OF CYCLES

Like the weather, cycles are a perennial topic of conversation, but as yet we are not in a position to do much about them. The term cycle refers to what the layman calls changes from "good" to "bad" business and back again. For the economist, the term cycle refers to the wavelike or undulatory movements of economic activity. Such questions as "What lies ahead next year for American business?" are questions mainly concerning cycles. The economy is not static, but has been marked by wide swings from prosperity to crisis to depression to recovery.

There is a vast area of disagreement among economists generally, and even among economists specializing in cycle theory, as to the characteristics of cycles. ". . . The fact that the two phenomena [boom and slump] are frequent and well known and possibly more studied than any other characteristic piece of economic behavior does not mean that there is anything like agreement on how they are caused or how to control them or how stability can be maintained in an economy." * Many involved theories have been and are being pro-

* Barbara Ward, *Policy for the West*, p. 118. W. W. Norton & Company, Inc., 1951.

pounded concerning the causes of cycles, their duration, the signs by which they manifest themselves (sometimes called "indicators"), and the ensuing effects.

Since the phenomenon cannot be precisely defined in its qualities, precision in quantitative terms is accordingly not to be expected. The utmost exactitude in mathematics does not help in overcoming this obstacle.

But the problem is so important that we must do the best we can in providing quantitative information. Even approximations that the statistician can supply will be valuable to the economist, the businessman, the government, and the citizen, providing the statistician's findings are seen against the background of inherent limitations. The alternatives to these limited statistical findings would be guesswork, hunches, and crystal gazing. And the fact that many excellent minds are devoted to studying this important question gives us ground for hope that by persistent, organized statistical analysis we shall continually advance toward an answer to the why and wherefore of cyclical fluctuations.

Finally, we must note, in passing to the statistical problems involved in cyclical variation, that cycles are thought to be related to the economic position of a country, and consequently vary from national economy to national economy.

Statistical Characteristics of Cycles

Cycles are not fluctuations which repeat themselves with periodic regularity, as do seasonal fluctuations. But neither are they fortuitous and haphazard like irregular fluctuations. They are in an intermediate position.

There appears to be a family resemblance between different cycles, in duration and intensity. Certain broad patterns do recur, but with no apparent regularity. For example, cyclical movements in steel production, for the last fifty years, show some similarity, but there is no exact repetition with regard either to the duration of the cycle or to its intensity.

This similarity of pattern is the basis for the ambitious at-

tempt of the National Bureau of Economic Research to arrive through statistical analysis at a pattern followed by cyclical movements. The first step here is to establish a large number of cyclical movements of a particular series (such as one company, or one industry, or an entire national economy). The second step is to arrange every cycle into nine stages. Then an average is obtained for each stage. These nine averages are thought to be the typical course of a cycle, or the cyclical pattern. This description, of course, is an oversimplified summary of the very involved procedure followed by the National Bureau of Economic Research.

What the National Bureau method seeks to find is a general law of business cycles through study of the interrelated fluctuations of many specific cycles, such as cyclical variation in steel production, in the movement of freight, in building construction, in employment, in interest rates, in profits, in the level of wholesale prices, in imports and exports, in the volume of savings, in bank debits, and in other series. These elements of the economy, according to the National Bureau of Economic Research, expand and contract at different rates and do not necessarily all expand at one and the same time in an upswing nor do they all contract at one and the same time in a downswing.

One outcome of the large-scale statistical undertaking of the National Bureau of Economic Research is that it may afford insight into the future course of the cyclical development in each series studied, and comparison of the timing of specific cycles. Another is that any specific cycle may be compared with the business cycle of the country as the frame of reference.

In basic statistics an approach much less involved than that of the National Bureau of Economic Research must suffice.

We already know that time series are thought to be shaped by a composite force symbolized by *TSCI*. If we can eliminate trend and seasonals, and perhaps irregular factors, we will be left with cycles. To be sure, complete elimination is not possible and the results obtained through eliminating *TSI* will

give us an approximation of cyclical movements but not an exact measurement of them. In this method C is left as a residue, as it were, and this method of approximating cycles is called *the residual method*.

MEASURING CYCLES BY THE RESIDUAL METHOD

Annual Data

Annual data are usually influenced by only two elements of the composite force, TC . Seasonal variations do not show up in annual data. Irregular movements, which are ordinarily of short duration (compared to a year), usually have small effect; irregular up and down movements tend to offset each other during the course of a year. In annual data, therefore, we can usually disregard the influence of irregular factors without distortion.

Thus when studying annual data we are left with the necessity of eliminating trend only. Hence we obtain an estimate of cyclical variation by dividing annual data for each year by the trend value for that year.

This elimination process involves a comparison of the original data with the so-called *statistical normal*. "Normal" here refers to what, for example, an entrepreneur considers to be the expected annual business. He calls business "normal" if it complies with the growth factor of his enterprise and is thus not influenced by what he may call abnormal factors, namely, cyclical fluctuations and irregular interferences. Thus, *in annual data the statistical normal is trend*.

The procedure in getting at cyclical values for annual data thus involves "adjustment for trend."

Table 17.1 shows the cyclical relatives for 1934 to 1940 obtained by dividing the original data by the trend values for all private corporate profits in the United States before taxes. Chart 17.1 shows the original data and the least-squares trend line, and Chart 17.2 the cyclical relatives.

TABLE 17.1. CORPORATE PROFITS BEFORE FEDERAL AND STATE INCOME AND EXCESS-PROFITS TAXES FOR ALL UNITED STATES PRIVATE INDUSTRIES, 1934-1940.

Year	Profits in Billions of Dollars, TC^*	Trend values, $T^\dagger$	Cyclical relatives, C^*
1934	1.6	2.2	73%
1935	3.1	3.1	100
1936	5.6	4.1	137
1937	6.1	5.0	122
1938	3.2	6.0	53
1939	6.4	7.0	91
1940	9.2	7.9	116

* The presence of I is disregarded here.

† Although the annual increment is always the same, the differences between these T values are not exactly the same, due to rounding.

Source: United States Department of Commerce, Bureau of Foreign and Domestic Commerce, 1948.

Billions of dollars

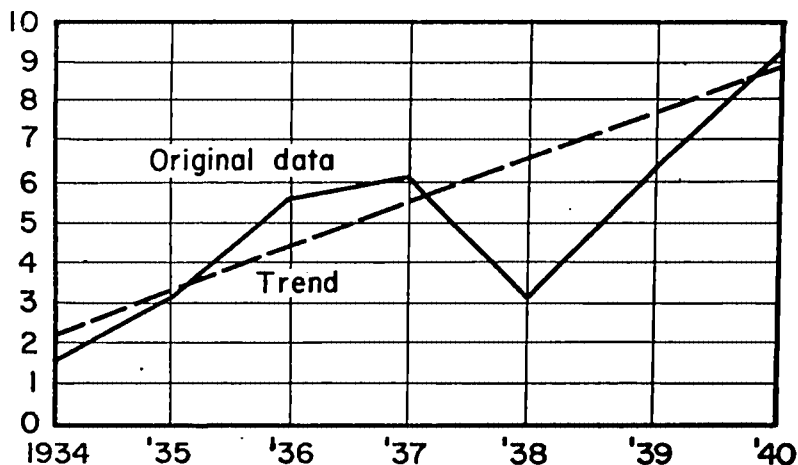


Chart 17.1. Original Data and Trend Line for Corporate Profits before Federal and State Income and Excess Profits Taxes for All United States Private Industries, 1934-1940.

Source: Same as Table 17.1.

Percent of trend

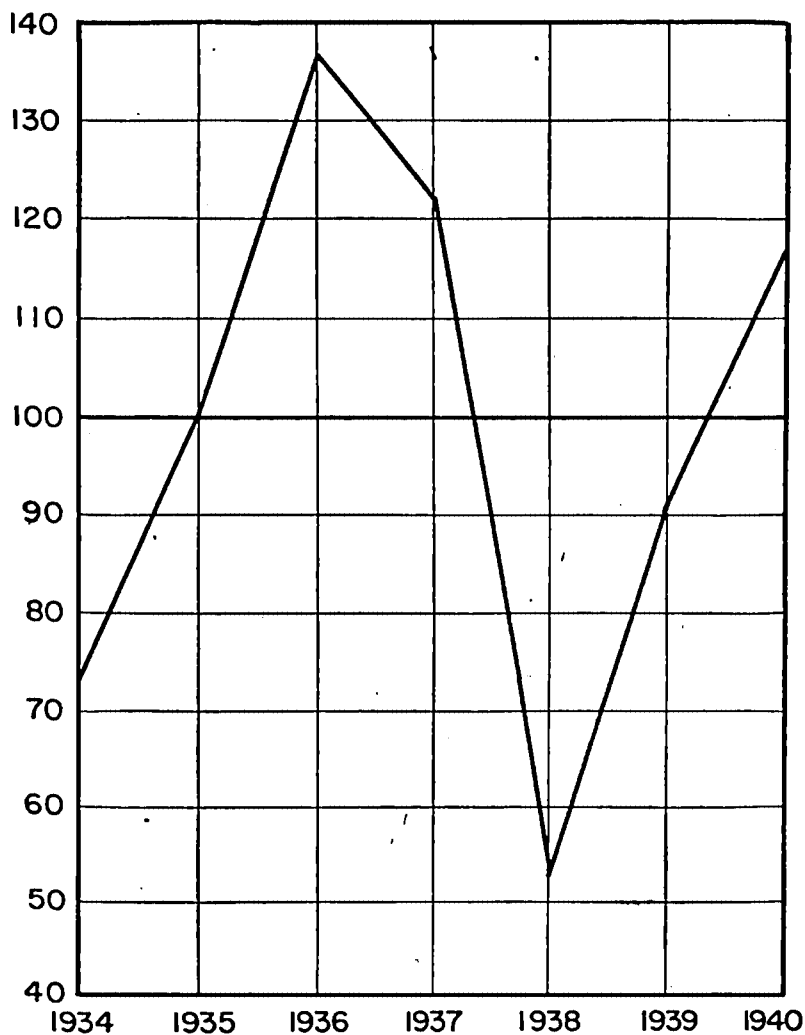


Chart 17.2. Cyclical Relatives for Corporate Profits before Federal and State Income and Excess-Profits Taxes for All United States Private Industries, 1934-1940.

Source: Table 17.1.

Monthly Data

In monthly data we find all four elements of the composite force, that is, *TSCI*. Therefore we must remove trend and deseasonalize the data, as well as eliminate the influence of irregular variations. Since irregular variations are short-lived, they usually will show up in monthly data.

For monthly data, in obtaining cyclical relatives we compare the original data with the statistical normal as we did for annual data. But *in monthly data the statistical normal consists not only of the growth factor as in annual data but also of the seasonal factor*. Thus, the expected monthly business is a combination of trend and seasonality, or *TS*. Accordingly

$$\frac{TSCI}{TS} = CI.$$

CI symbolizes the cyclical-irregular relatives, and we must seek to iron out the irregular element.

The procedure for obtaining the cyclical relatives for monthly data involves the following three steps:

1. Divide the original data for each month by the trend value for this month, that is, compute $TSCI/T = SCI$.
2. Divide the monthly data adjusted for trend by the seasonal index for this month, that is, compute $SCI/S = CI$. We thus obtain the cyclical-irregular relatives.

The sequence of steps 1 and 2 may be reversed.

3. We now have *CI* values expressed in percent. Since we are dealing with monthly data, we must seek to remove *I*, which we do by subjecting the *CI* values to the ironing-out process of a moving average. The appropriate period of the moving average depends on the average duration of the irregular variations. For instance, a three-month moving average may be used if the irregular variations average about three months in time.

A second and shorter method for isolating cyclical variations from monthly data makes use of the computational work in

TABLE 17.2. CYCLICAL RELATIVES FOR PRODUCTION OF PORTLAND CEMENT IN THE UNITED STATES, JANUARY 1937 TO DECEMBER 1939.

		Production, millions of barrels, <i>TSCI</i>	Centered 12- month moving average <i>TC</i>	Trend value, <i>T*</i>	Cyclical relative, <i>C</i>
1937	J	6.6	10.25	8.68	118.1
	F	5.8	10.22	8.76	116.7
	M	8.4	10.15	8.84	114.8
	A	10.4	10.05	8.92	112.7
	M	11.6	9.93	9.00	110.3
	J	11.2	9.78	9.08	107.7
	J	11.6	9.60	9.16	104.8
	A	11.9	9.44	9.24	102.2
	S	11.2	9.25	9.32	99.2
	O	11.4	9.05	9.40	96.3
	N	9.2	8.90	9.48	93.9
	D	7.0	8.82	9.56	92.3
1938	J	4.5	8.77	9.64	91.0
	F	3.9	8.70	9.72	89.5
	M	5.9	8.64	9.80	88.2
	A	8.0	8.63	9.88	87.3
	M	10.4	8.68	9.96	87.1
	J	10.5	8.76	10.04	87.3
	J	11.0	8.84	10.12	87.4
	A	11.0	8.94	10.20	87.6
	S	10.6	9.10	10.28	88.5
	O	11.6	9.27	10.36	89.5
	N	10.2	9.38	10.44	89.8
	D	8.1	9.47	10.52	90.0
1939	J	5.3	9.60	10.60	90.6
	F	5.5	9.73	10.68	91.1
	M	8.2	9.84	10.76	91.4
	A	9.7	9.93	10.84	91.6
	M	11.2	10.00	10.92	91.6
	J	12.0	10.10	11.00	91.8
	J	12.6	10.20	11.08	92.1
	A	12.4	10.21	11.16	91.5
	S	11.9	10.18	11.24	90.6
	O	12.5	10.18	11.32	89.9
	N	11.1	10.25	11.40	89.9
	D	9.5	10.33	11.48	90.0

* The trend equation fitted to monthly data from January 1933 to December 1942 is

$$Y_c = 9.64 + .08X$$

origin, January 1938;

X units, one month;

Y units, production in millions of barrels.

Source: *Survey of Current Business*, various issues.

obtaining the seasonal index. There (as can be seen on page 275), we found that the centered 12-month moving average is an estimate of TC . By dividing the 12-month moving average for each month by the trend value for this month, or computing TC/T , we obtain an estimate of C . Since the 12-month moving average minimizes the influence of irregular factors, there is no need to adjust for them.

Table 17.2 illustrates the procedure for obtaining cyclical relatives by eliminating trend from the centered 12-month moving average for the production of Portland cement in the United States for the years 1937–1939. In Chart 17.3, these cyclical relatives are plotted.

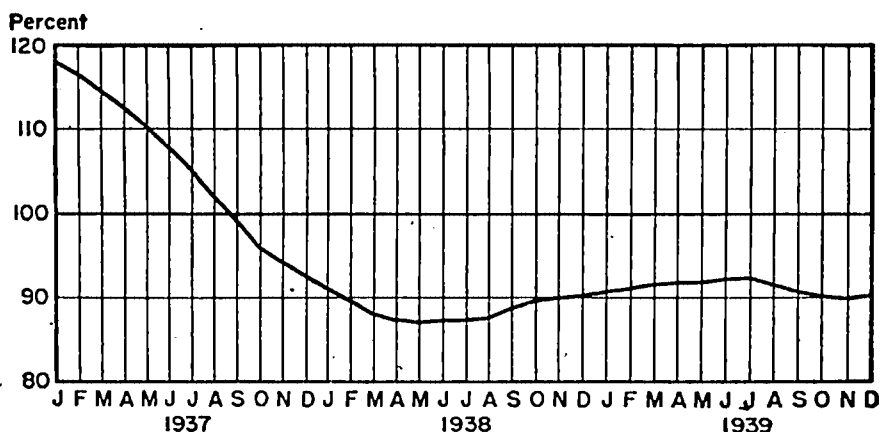


Chart 17.3. Cyclical Relatives for Production of Portland Cement in the United States, January 1937 to December 1939.

Source: Same as Table 17.2.

The procedure for obtaining cyclical relatives for quarterly data is analogous to that for monthly data. Of course, in quarterly data the TC values are represented by a centered 4-quarter moving average.

IRREGULAR FACTORS

As a rule, we are interested only in *eliminating* irregular factors; by themselves they usually do not have intrinsic im-

portance. But if we wish to study irregular factors by themselves in monthly data, then we can revert to a residual method. For example, if we wish to study fluctuations due to irregular factors in steel production in two or more different periods (for instance, when there were coal strikes), we would seek to isolate the irregular factor for each period. Since we can successively find T , S , and C , we can step by step remove each of them from the original data $TSCI$, and obtain I . But we are not able to precisely eliminate T , S , and C , and therefore there are impurities in the final residue. Hence, the estimates of the irregular factor are very crude approximations.

A shorter method for isolating the irregular factor is to use the seasonal relatives SI and eliminate the seasonal index S by dividing it into SI . This method was used in separating the irregular variations from the seasonal relatives for factory production of creamery butter for the years 1943-1947, which were plotted in Chart 14.7.

FORECASTING

Importance of Forecasting

In everyday speech, forecasting means "looking into the future." In statistics, the term refers to extending or projecting time series into the future* based on past behavior of the quantitative data. Economics, business, and government cannot wait for the future to overtake them; they must plan for it. All three look for objective guides to what is going to occur.

Forecasts in economics and business must be made for purposes such as: predicting the future course of the whole economy; judging future markets; gauging future employment; making decisions on production, buying, inventories, advertising, selling, and pricing. The aim of forecasting is to establish, as accurately as possible, the probable behavior of economic activity based on all data available, and to set policies in terms

* Looking backward into a past, for which data are not available, statistically is analogous to forecasting.

of these probabilities.* Statistical forecasting has reached a point where, even though it is necessarily limited, it does give us a basis in fact from which to make decisions.

Methods of Forecasting

Broadly, there are three ways of forecasting: (1) rule-of-thumb method; (2) qualitative approach; (3) statistical method.

1. The rule-of-thumb method is widely practised; it consists of deciding about the future in terms of past experience and familiarity with the problem at hand. To be sure, this method can lead to absurd conclusions if employed by the inexperienced, but many businessmen have used it successfully on a small scale.

2. In the qualitative approach, we base decisions about the future on nonstatistical data and arguments. These data and arguments may be historical, political, psychological, economic. For example, after World War II predictions were made by historical analogy to the economic situation following World War I. Forecasts are sometimes based mainly on anticipations in national and international politics. In general, the qualitative approach involves predictions on the basis of what may be called the "unmeasurable" variables.

3. Statistical forecasting can be done in terms of three of the four elements of the composite force. These three are trend, seasonality, and cycles. We attempt to predict the continuance of some tendency or the recurrence of some pattern. Statistical forecasting cannot, obviously, embrace irregular variations since in their very nature they are unpredictable.

Procedures in Statistical Forecasting

In *trend forecasting* we project the trend line into the future, either graphically or by substituting in the trend equation the X value corresponding to the future date. This method is known as *extrapolation*. Thus we project into the future the growth

* See Leo Barnes, *Handbook for Business Forecasting*, 2nd revised edition, Prentice-Hall, Inc., 1949.

factor analyzed in the past. This projection is really an extension of the statistical normal if we are dealing with annual data.

But in monthly or quarterly data the statistical normal is trend combined with *seasonality*. Here, therefore, we forecast by finding the future monthly or quarterly trend value from the trend equation and then multiplying it by the value of the seasonal index for this month or quarter. Thus we project the statistical normal (*TS*) for monthly or quarterly data into the future.

There is no difficulty in forecasting seasonality alone; the seasonal index is in itself a forecast.

Cyclical forecasting presents thorny problems. Forecasting can be done only if there is believed to be something resembling a cyclical pattern. But even though each business cycle is unique, there are among cycles some family resemblances which induce some statisticians to see something of a pattern. Based on where present business conditions fall in this pattern, these statisticians forecast the next cyclical development. We have already mentioned the work of the National Bureau of Economic Research which aims at a "cyclical pattern."

If we were able to predict cyclical values, then we could combine these predictions for given data with trend-seasonal forecasts. The result would be a "complete forecast" except for the unpredictable irregular variations.

Limitations of Statistical Forecasting

It has been said that the only thing certain about the future is that it is uncertain. This limitation holds for statistical forecasting. As one author has put it: "The only thing you can be sure about in any forecast is that it will contain *some* error." *

Statistical forecasting is done under the assumption "other things being equal," which means that the conditions operating in the data remain the same. But they rarely do. In long-term trend, we give equal importance to the comparatively distant past and the recent past, but the recent past may be

* Clarence Judd, "Do Your Own Forecasting," *Steelways*, January 1950.

more indicative of future developments. Moreover, prediction of a trend value alone may be very wrong if the time period for which the prediction is made turns out to be at the high or low of a cycle.

If the trend is downward then extrapolation may give us a negative figure. If we mechanically extend the trend line for the declining death rate from tuberculosis into the future, we may get a negative death rate, obviously preposterous, since what is happening is that the death rate from tuberculosis is leveling off.

Seasonal forecasting assumes an unchanging seasonal pattern. But, as we know, seasonal patterns have been known to change. Especially abrupt changes of the pattern destroy the reliability of a seasonal forecast.

The most stringent limitations attach to cyclical forecasting. The definition of a cycle is not generally agreed upon, and the existence of a cyclical pattern has been disputed.

These limitations of statistical forecasting require that use be made of any available information on the nonstatistical level. Thus, more successful forecasting will result, we think, by combining with statistical forecasting the "rule-of-thumb" method and the qualitative approach.

Summary

1. Cycles are very difficult to measure, but statistical measurement is the alternative to guesswork.
2. Cycles do not recur at regular periods, nor are they entirely irregular.
3. The residual method for arriving at cyclical relatives consists of isolating cycles from the other three elements of a time series.
4. In annual data, we eliminate trend from the original data and thus approximate cyclical relatives.

5. In monthly or quarterly data, we must eliminate trend, seasonality, and irregular variations in order to estimate the cyclical relatives.

6. Irregular variations may be estimated by the residual method also.

7. Forecasting means predicting future values of trend, seasonality, and cycles.

8. Despite great limitations in statistical forecasting, it is invaluable as a guide in the planning of the economist, the businessman, and the government.

CHAPTER

18

Index Numbers

Importance of Index Numbers *

In industrial relations such situations as the following have frequently occurred. The labor union representing the workers in a large industry was in collective bargaining with a major corporation in that industry. The chief stumbling block in the negotiations was the problem of wages. Since the cost of living had been rising for some time, the union's representatives were wary of accepting a simple blanket wage increase for the duration of the contract, which might extend over several years. Therefore they offered to accept a wage increase of a given amount with the proviso in the agreement that if the cost of living thereafter should rise by a certain amount, the wages were to rise proportionally. The plan was accepted by the corporation with the proviso that if the cost of living should drop by a certain amount wages were to be lowered accordingly.

How do we measure changes in the cost of living, or changes in the prices that consumers pay? The measuring rod is statistical, and makes it possible to compare changes in groups of prices from time to time or from place to place. It is called an *index number*. In addition to measuring changes in consumers'

* The terms *index* and *index number* are in practice used interchangeably, although rigorously speaking an *index* might be restricted to mean a *series* of index numbers.

prices, index numbers also measure changes in wholesale prices, industrial production, sales—in short, any variable capable of showing changes from time to time and from place to place.

Index numbers are today one of the most widely used statistical devices. Newspapers headline the fact that prices are going up or down, that industrial production is rising or falling, that sales are higher or lower than in a previous period, as disclosed by index numbers. They are used to take the pulse of the economy, and they have come to be used as indicators of inflationary or deflationary tendencies. In time-series analysis, index numbers are used to adjust the original data for price changes, or to adjust wage changes for cost-of-living changes and thus transform nominal wages into real wages. Moreover, nominal income can be transformed into real income, and nominal sales into real sales, through appropriate index numbers.

Index numbers are also used in educational, psychological, and sociological statistics.

Index Numbers and Other Statistical Concepts

In general, index numbers refer to groups of variables, such as the prices of building materials or of grains, the quantity of fuels consumed or of textiles produced. In all of these instances, we have a number of series each gathered into the same "basket" or group of commodities; for example, an index of fuel consumption may refer to the quantities of coal, fuel oil, gas, and kerosene consumed in one or more time periods. This type of index is usually referred to as a *composite index*, since it is the resultant of changes taking place in more than one series.

As distinct from a composite index, there are indexes which permit comparison only within one series, for instance the "index" of the number of business failures in the United States as compiled by Dun and Bradstreet. Such a single-series index, it is often claimed, is not really an index but a single series of percentages. These single-series indexes serve only to simplify the figures. From now on, when we speak of index numbers, we refer only to composite indexes.

In order to arrive at an index number, we must represent a number of values by a typical summary figure. Thus the concept of the average enters into the construction of index numbers. Some index numbers may indeed be considered as averages of a specialized type. They thus have the limitations of averages. An average, it will be recalled, does not tell the full story but must be supplemented by a measure of dispersion. In index numbers, too, extreme values may distort the mean, and there may be wide dispersion, making any average a poor representative value. For example, in a given "basket" of commodities, prices of one commodity may rise sharply in one year compared to another, while prices of another commodity may decline, and the index number consequently may show no change. Thus, the average level remains the same, not giving any indication of the changes in dispersion. (Desirable as it may be to use measures of dispersion relative to index numbers, this practice has not become established.)

Moreover, most frequently, the items that enter into the construction of index numbers are not exhaustive. We usually take a sample* of the items involved in our problem, not the universe. For instance, the Wholesale Price Index for the United States, compiled by the Bureau of Labor Statistics, does not include all wholesale prices in the United States, which would be the universe, but draws conclusions from a sample consisting in 1951 of about 900 commodities. The sample required for constructing index numbers is generally purposive, and sometimes stratified as well. Thus, usually, the requirements and limitations of sampling must be kept in mind in the construction and use of index numbers.

Index numbers usually are comparisons over time rather than place. Thus they are time series. In certain cases, index numbers partake of time series to such an extent that aspects of time-series analysis enter into their construction and use.

* *Sample*, *universe*, *purposive*, and *stratified* are technical words which are defined and explained in Part 4 of this book, particularly in Chapter 21, pages 381 ff., and Chapter 23, pages 430 ff.

With relatively few exceptions, index numbers are expressed in percent, and thus the concept of ratio is also important here.

Classification of Index Numbers

Index numbers may be classified in terms of what they measure. In economics and business the classifications are (1) price; (2) quantity; (3) value; (4) special purpose.

Price indexes are illustrated by the Wholesale Price Index for the United States of the Bureau of Labor Statistics; quantity index numbers by the Index of Industrial Production of the United States of the Federal Reserve Board; value index numbers by the Index of Department Store Sales, also computed by the Federal Reserve Board; and special-purpose index numbers by the *New York Times* Index of Business Activity.

We need go into the details of constructing only price and quantity index numbers. The others will be mentioned, but without details of how to construct them since both value and special-purpose index numbers do not offer new problems in construction. Since the details of construction of all types of index numbers can be understood if the construction of price index numbers is understood, we shall devote major attention to them.

PROBLEMS IN THE CONSTRUCTION OF PRICE INDEX NUMBERS

It is absolutely necessary that the purpose of the index number be rigorously defined. Thus, a price index that is intended to measure consumers' prices must not include wholesale prices. And if such an index is intended to measure the cost-of-living of moderate-income families, great care should be exercised not to include goods ordinarily used by upper-income groups. As everywhere else in statistics, we must know what we want to measure and what we intend to use the measurements for.

Data

The problems of data are especially acute in constructing index numbers. A large number of factual questions is involved in amassing the data. The variety of goods and prices makes the selection of data a prime consideration. Ordinarily, we draw upon many sources of data which are geographically dispersed. Problems of comparability and reliability thus multiply and the chances for spurious results are increased. One mistake may "bias" the index. Examples of such a mistake would be: including the price of one commodity in the "basket" for one time period and the price of a slightly different commodity in the "basket" for another period; or taking the manufacturer's price of some commodity at one time, the wholesaler's or jobber's price at another.

We must decide at the very inception of the inquiry whether we are going to collect the data ourselves or rely on a published source. The labor involved in the collection of this kind of data is of vast extent. Moreover, collection of index-number data is not a one-time task. Index numbers of prices are ordinarily computed monthly or quarterly. In some instances, they are found weekly and even daily.

In most cases, it will not be feasible to collect data on the universe of prices, and a sample is called for. The selection of the sample involves careful consideration; for example, prices of obsolescent types of clothing should not be represented in a clothing-price index. If such items are, or continue to be, included, the index is distorted.

Base

In order to make comparisons between prices referring to several time periods or several places, some point of reference is almost always established. This point of reference is called the base. The prices at a certain time period (or place) are taken as the standard, and to them is assigned the value of 100%. There are two important guide lines to consider in choosing a base.

1. **THE BASE AS TYPICAL.** If we take as a base the prices in a time period of prosperity, then prices in other time periods look low. On the other hand, if we choose as a base prices in a depression period, prices in all other periods look high. Therefore, we seek as a base the prices in a time period that conforms with trend rather than a period with high or low cyclical values. That is, we choose a normal period, and the difficulties in defining "normal" are as great here as anywhere else. But what we can do is avoid extremes which lead to distortions.

Sometimes it is difficult to choose just one year as the normal. In such cases, taking the average of a few years will reduce the effect of extremes. Thus, the average of the whole period from 1947 to 1949 may be considered normal, whereas no individual year in that span can be considered normal.

2. **CHOICE OF A BASE NOT TOO FAR IN THE PAST.*** *Choosing a Base.* Since practical decisions are made in terms of index numbers, and economic practices so often are a matter of the short run, we wish to make comparisons between a base which lies in the same general economic framework as the years of immediate interest. Therefore, we choose a base relatively close to the years being studied.

Changing a Base. If new items are added to the data, then the base must be established at a period when these new items influence the index. Moreover, if the data at an early period are no longer considered reliable in terms of new statistical criteria, then the base should be shifted to a period when prices were collected on the more reliable foundation.

In both cases, index numbers constructed on the new base are not wholly comparable to index numbers referring to any time preceding the new base period.

Combining the Data

We are interested, let us suppose, in establishing a price index of leading heating fuels in a city on the eastern seaboard.

* For technical considerations bearing on this topic, see Frederick C. Mills, *Statistical Methods*, Third Edition, Henry Holt and Company, 1955, pp. 470-471.

We start with the prices of coal, fuel oil, gas, and kerosene for each year from 1952 to 1955. The problem is to combine price data for each year into one expression.

This combining of prices for each year can be done either by totaling or by averaging. Totaling the prices for each year leads to (1) the *simple aggregate of actual prices* for each year. Averaging leads to (2) the *simple average of price relatives*.

1. Simple Aggregate of Actual Prices

For each year, we total the prices of the items in the basket of goods—in our example in Table 18.1, the wholesale prices of dairy products. The aggregate cost of the basket in a year is compared with the cost of this basket in other years. The comparison may be presented simply in dollars and cents. More often and more usefully, it is presented in percent, the aggregate cost for some year being taken as 100%. This 100% year is the *base year*.

TABLE 18.1. INDEX OF WHOLESALE PRICES OF DAIRY PRODUCTS IN THE UNITED STATES FOR 1949, 1950, AND 1952, COMPUTED BY THE METHOD OF SIMPLE AGGREGATE OF ACTUAL PRICES (1949 = 100).

Commodity and unit	Price, 1949, p_o	Price, 1950, p_n	Price, 1952, p_n
Creamery butter, lb.	.615	.622	.730
American cheese, lb.	.348	.354	.441
Condensed milk, case.	9.17	9.25	10.80
Fluid milk, 100 lb.	4.76	4.57	5.46
Eggs, doz.	.500	.420	.468
Total.	\$ 15.393	\$15.216	\$ 17.899
$\frac{\Sigma p_n}{\Sigma p_o}$		$\frac{15.216}{15.393}$	$\frac{17.899}{15.393}$
Index, P	100.0	98.9	116.3

Source: *Survey of Current Business*, 1953 Supplement, titled *Business Statistics*, 1953 Biennial Edition, United States Department of Commerce.

In Table 18.1 we have selected as a base the year 1949. This is expressed as follows: $1949 = 100$. Here the aggregate of actual prices for 1949 for dairy products is \$15.393. The aggregate of the actual prices for 1950 is \$15.216. The 1950 aggregate is compared with the aggregate of actual prices in the base year. Dividing the 1950 aggregate by the base-year aggregate gives us an index for 1950 with 1949 as a base. For 1950 a decrease of 1.1% is indicated over the base period. The same procedure is followed for the other year in the series. The results are shown in Table 18.1.

Each price of each commodity is designated by the symbol p ; p_o stands for the price of a commodity in the base period, p_n stands for the price of a commodity in any other period but the base period, and such a period is called the *given* period. The capital letter P stands for the price index. The aggregate of actual prices in the base year is Σp_o . The aggregate of actual prices in a given year is Σp_n . Therefore, the formula for finding a simple aggregate of actual prices for each time period in a series is

$$P = \frac{\Sigma p_n}{\Sigma p_o}$$

In the example of Table 18.1 we had \$15.393 for Σp_o , \$15.216 for Σp_n , and \$17.899 for the other Σp_n . In Table 18.1 we thus got an index number of 98.9% for 1950 and an index number of 116.3% for 1952.

Let us summarize the steps:

1. Total the prices of the various commodities for each time period to get Σp_o and Σp_n . These totals are in dollars.
2. Divide the total of the given time period, Σp_n , by the base-period total, Σp_o . This result is expressed in percent.

2. Simple Average of Price Relatives

As the name of this type of index implies, it consists of finding price relatives and then averaging them.

First, a "price relative" is obtained by expressing the price

of each commodity in a given year as a percent of the price of the commodity in the base year. Thus, we divide the price of butter in 1950 by the price of butter in 1949, and express this in percent. Therefore, any single price relative is symbolized by p_n/p_o .

Then we have to average the price relatives for every time period. Any average may be used. Theoretically, the geometric mean is best in averaging percentages or rates. (See Chapter 11, page 160.) But in practical statistical work, the arithmetic mean and the median are widely used for reasons of simplicity. We will here use the arithmetic mean.

A simple average of price relatives for dairy products with 1949 as a base is shown in Table 18.2. For example, the simple average of price relatives for 1952 is found by adding 118.7%, 126.7%, 117.8%, 114.7%, and 93.6%, then dividing the sum (571.5%) by the number of items, which is 5. This arithmetic

TABLE 18.2. INDEX OF WHOLESALE PRICES OF DAIRY PRODUCTS IN THE UNITED STATES FOR 1949, 1950, AND 1952, COMPUTED BY THE METHOD OF SIMPLE AVERAGE OF PRICE RELATIVES (1949 = 100).

Commodity and unit	Price			Price relative		
	1949, p_o	1950, p_n	1952, p_n	1949, p_o/p_o	1950, p_n/p_o	1952, p_n/p_o
Creamery butter, lb.	.615	.622	.730	100.0	101.1	118.7
American cheese, lb.	.348	.354	.441	100.0	101.7	126.7
Condensed milk, case	9.17	9.25	10.80	100.0	100.9	117.8
Fluid milk, 100 lb.	4.76	4.57	5.46	100.0	96.0	114.7
Eggs, doz.	.500	.420	.468	100.0	84.0	93.6
Total.				500.0	483.7	571.5
$\frac{\sum \frac{p_n}{p_o}}{N}$					$\frac{483.7}{5}$	$\frac{571.5}{5}$
Index.				100.0	96.7	114.3

Source: Same as Table 18.1.

mean gives an index number of 114.3 for 1952 with 1949 as the base.

Or to summarize: (1) We get the price relative by dividing the price of each commodity in the given time period, p_n , by its price in the base period, p_o , and expressing this result in percent. (2) We then average these relatives for the given time period. The formula, when the arithmetic mean is used, is

$$P = \frac{\sum \left(\frac{p_n}{p_o} \right)}{N}$$

If other averages are used, the formula is of course different.

3. Shortcomings of the Simple Aggregate of Actual Prices and the Simple Average of Price Relatives

In the simple aggregate of actual prices in Table 18.1, the price of condensed milk dominates the index. It is expressed per case and is much the largest component of the aggregate price of these dairy products. We can readily see that if the price of condensed milk were expressed per pound the index would be very different. A 10% change in the price of condensed milk in one direction has a much stronger influence on the index than a 10% price change in the other direction for the four other dairy products combined. This tremendous influence of the condensed-milk price on the index results not necessarily from its economic importance, but solely from the fact that the price of condensed milk is quoted per case. Thus, the unit by which each item happens to be priced introduces a concealed weight in the simple aggregate of actual prices. This concealment is undesirable and severely restricts the usefulness of an index number arrived at through the method of the simple aggregate of actual prices.

The attempt has been made to express product prices in the same unit in the simple aggregate of actual prices; for instance, to express each commodity price per pound. But this practice does not obviate the difficulty. First, in a cost-of-living index, for

example, services cannot be expressed in pounds. Furthermore, even with the same pricing unit, an undesirable emphasis on one commodity may appear; for example, in a food-price index the price of one pound of tea is many times the price of one pound of potatoes.

In the simple average of price relatives, this difficulty does not appear, because we are comparing price per pound with price per pound and price per ton with price per ton; in condensed milk, for example, the price of \$10.80 per case in 1952 is compared with the price of \$9.17 in 1949, giving a price relative of 117.8% as shown in Table 18.2. Nevertheless, we find a concealed assumption. The concealed assumption is that in the base year (in the data in Table 18.2) the wholesalers sell an equal number of dollars' worth of each commodity in the basket. This assumption does not correspond to experience. Let us examine this assumption as if \$100 worth of each commodity were sold in the base year. Thus in 1949, as set forth in Table 18.2, the wholesalers sold \$100 worth of creamery butter. For this amount of money in the base year 1949 they marketed

$$\frac{\$100}{\$.615} = 162.6 \text{ pounds.}$$

Similarly as to eggs; for \$100 they sold

$$\frac{\$100}{\$.500} = 200 \text{ dozen eggs.}$$

Now consider the year 1952. How much do the wholesalers charge for the same quantities in 1952? The butter price in 1952 is \$.730; therefore they get \$118.7 for 162.6 pounds of butter. But for 200 dozen eggs they get \$93.6. These last two dollar figures, \$118.7 and \$93.6, are the same as the price-relative figures for the year 1952, because we have carried over into the given year 1952 the assumption that the wholesalers sell 162.6 pounds of butter and 200 dozen eggs. That is to say, we have embedded these quantities in the index; in other words, we have

unintentionally weighted the data. Since, as was mentioned, the assumption that gave us these weights does not correspond to experience, the weighting is undesirable.

In both methods—the simple aggregate of actual prices and the simple average of price relatives—concealed and usually undesirable weights are present. Better index numbers will be obtained if we bring out the relative importance of the commodities by openly applying appropriate emphasis. This open application of appropriate emphasis is known as *weighting*.

Weighting

In order to apply appropriate weights in a price index, we must answer the following questions: (1) By what do we weight? (2) What type of weight do we use? (3) From what time period do we take the weights?

1. We can weight by whatever seems appropriate to bring out the economic importance of the commodities involved. The weight can be production figures, consumption figures, or distribution figures. The statistician here makes a decision based upon economic knowledge.

2. There are two types of weights: quantity weights and value weights. A quantity weight, symbolized by q , means the amount of a commodity produced, distributed, or consumed in some time period. A value weight combines price with quantity produced, distributed, or consumed. Value means dollar volume and is symbolized by $p \times q$.

The statistician is not free to choose here. If we use the method of aggregates, then quantities can be used as weights, because price times quantity will always give the same units, namely dollars. But in the case of price relatives we cannot use quantity figures. If we multiply percentages by quantities expressed in different units, we get results in different units; for example, percentages times tons will give tons and percentages times pounds will give pounds. Such figures cannot be used in computation. But if we multiply percentages by value figures, which

are always expressed in dollars, we get answers in dollars only.

Therefore, the statistician will use q as a weight in the method of aggregating actual prices and must use $p \times q$ as a weight in the method of averaging price relatives.

3. As for the time from which we take the weights, let us first consider quantity weights. They may be taken from the base period of the index, symbolized by q_0 , or from the given period, symbolized by q_n , or as a sum or average of the two. Widely used as quantity weights are those taken from a period considered typical, symbolized by q_t . Sometimes, averages of quantities in more than one typical year are used; for example $q_{1947-49}$, which is used for part of the United States Wholesale Price Index.

As for value weights, a combination of p in any time period and q in any time period may be chosen. But in practice, base-year values are used most frequently; that is, p_0q_0 .

Thus, we may apply weights to the prices that enter into the simple aggregate of actual prices and arrive at the weighted aggregate of actual prices; we may likewise apply weights to the price relatives that enter into the simple average of price relatives and arrive at the weighted average of price relatives.

1. Weighted Aggregate of Actual Prices

Let us find the weighted aggregate of actual prices for dairy products in 1949, 1950, and 1952, with 1949 as a base, using base-year weights. These weights consist of quantities produced in 1949. Weights are always applied by multiplication.

The index number for each year is obtained in three steps:

1. The price of each commodity in each year is multiplied by the base-year quantity of that commodity. For the base year each product is symbolized by p_0q_0 , and for each given year by p_nq_0 .
2. We total the products for each year. This gives us Σp_nq_0 for each given year and Σp_0q_0 for the base year 1949.
3. We divide each given-year total by the base-year total.

Thus, the base-year weighted aggregate of actual prices* is symbolized by the formula

$$P = \frac{\sum p_n q_o}{\sum p_o q_o}$$

The computation of this index is shown in Table 18.3.

2. Given-Year Weights

Instead of base-year weights, given-year weights (q_n) may be used. In so doing, for a 1950 index number, we must use 1950 weights; for 1952, we must use 1952 weights.

In using given-year weights certain new problems arise.

First of all, since the weights change with each given year, the base-year figures must be recomputed each time. A weighted index for the years 1949, 1950, and 1952, with 1949 as a base, using given-year weights, requires that 1949 prices be multiplied by 1950 quantities as the base for the 1950 index, by 1952 quantities as the base for the 1952 index. Far more work is thus required than in the use of base-year quantities.

Second, since the base-year figures are no longer stable, the index number for each given year is comparable only to that for the base year. Thus an index number for 1950 with base year 1949 is comparable to the index number for the base year 1949 but not to the index number for 1952, or any other year, when given-year weights are used.

Third, since the quantity weights of each given year are used, it becomes necessary to assemble quantity data for each given year. This task may be time consuming and it may be difficult to secure all the quantity data necessary.

The formula for the given-year weighted aggregate of actual prices† is

$$P = \frac{\sum p_n q_n}{\sum p_o q_n}$$

* This is known as Laspeyres' method, after Etienne Laspeyres, the German statistician and economist.

† This is known as Paasche's method, after Hermann Paasche, the German statistician and economist.

TABLE 18.3. INDEX OF WHOLESALE PRICES OF DAIRY PRODUCTS IN THE UNITED STATES FOR 1949, 1950, AND 1952, COMPUTED BY THE METHOD OF BASE-YEAR WEIGHTED AGGREGATE OF ACTUAL PRICES (1949 = 100).

* Commodity and unit	Price			Quantity* produced in 1949, millions of units, q_0	Price $\times$ weight		
	1949, p_0	1950, p_n	1952, p_n		1949, $p_0 q_0$	1950, $p_n q_0$	1952, $p_n q_0$
Creamery butter, lb.....	.615	.622	.730	117.7	72.39	73.21	85.92
American cheese, lb.....	.348	.354	.441	77.9	27.11	27.58	34.35
Condensed milk, case.....	9.17	9.25	10.80	8.4	77.03	77.80	90.72
Fluid milk, 100 lb.....	4.76	4.57	5.46	96.8	460.77	442.38	528.53
Eggs, doz.....	.500	.420	.468	390.0	195.00	163.80	182.52
Total.....					832.30	784.67	922.04
					$\Sigma p_0 q_0$	$\Sigma p_n q_0$	$\Sigma p_n q_0$
$\frac{\Sigma p_n q_0}{\Sigma p_0 q_0}$					832.30	784.67	922.04
					832.30	832.30	832.30
Index.....					100.0	94.3	110.8

Source: *Survey of Current Business*, 1953 Supplement, titled *Business Statistics*, 1953 Biennial Edition. United States Department of Commerce.

* Monthly average.

The same difficulties involved in using given-year weights arise when combinations of weights are used which include given-year weights. An instance occurs when the total or average of base-year and given-year weights is used.

Because of these difficulties, given-year weights are rarely used.

3. Typical-Year Weights

As distinct from the rare use of given-year weights, typical-year weights are of practical significance and use. The two major indexes computed by the Bureau of Labor Statistics make use of this method. These indexes use typical-year quantities; for example, the Wholesale Price Index uses q for 1947 as typical for the vast majority of commodities.

The formula here is

$$P = \frac{\sum p_t q_t}{\sum p_0 q_t}$$

where q_t stands for the quantity in the typical period.

Why do we sometimes find that a year or combination of years different from the price base year is used for weighting? The criteria that determine the "normal" for prices and the "normal" for quantities produced, distributed, or consumed are not necessarily the same.* The selection of typical-year weights may occasionally be influenced by availability of weight data.

4. Weighted Average of Price Relatives

Since price relatives are in percentages, the weights used must be value weights, as we have pointed out. If we use the arithmetic mean as an average, the formula becomes that of a weighted arithmetic mean of price relatives. The weighted mean, as we know (Chapter 10, page 131), has the formula $\sum wX / \sum w$. Since w , the weight, is $p_0 q_0$ if base-year values are

* For detailed discussion of the problem of weighting, see Bruce D. Mudgett, *Index Numbers*, John Wiley and Sons, 1951.

used as weights, the formula for the weighted average of price relatives is:

$$P = \frac{\sum \left(\frac{p_n}{p_o} p_o q_o \right)}{\sum p_o q_o}$$

This formula holds only when the arithmetic mean is used as the average for the price relatives. Since it is a weighted arithmetic mean, we divide by the sum of the weights ($\sum p_o q_o$) rather than by simply the number of items (N) as in the simple average of price relatives.

The formula can be transformed by cancellation in the numerator into

$$P = \frac{\sum p_n q_o}{\sum p_o q_o}$$

This transformed formula is the same as that for the weighted aggregate of actual prices. Why then do we treat it as a separate method of constructing price index numbers? The transformation is based upon two conditions not necessarily always present: (1) that the arithmetic mean is being used; (2) that base-year values are used as the weights. If these two conditions are not present, then the original formula is different and not susceptible to transformation in this way. Moreover, sometimes index numbers are constructed by combining existing index numbers. Since the data in this case—index numbers—are given in the form of *relatives*, the untransformed formula involving relatives must be used, and the transformed formula is therefore not applicable.

It may be advisable to use the untransformed formula for still another reason, inherent in the average-of-relatives method in general. Because computing the average of price relatives requires us to compute the price relatives for each single commodity in the basket (as cheese), we develop series of price relatives (as for cheese) and these series can be studied and analyzed in and of themselves.

TABLE 18.4. INDEX OF WHOLESALE PRICES OF DAIRY PRODUCTS IN THE UNITED STATES FOR 1949, 1950, AND 1952, COMPUTED BY THE METHOD OF BASE-YEAR WEIGHTED AVERAGE OF PRICE RELATIVES (1949 = 100).

Commodity	Price relative			Value weight, 1949, $p_0 q_0$	Weighted price relative		
	1949, $\frac{p_0}{p_0}$	1950, $\frac{p_n}{p_0}$	1952, $\frac{p_n}{p_0}$		1949 $\frac{p_0}{p_0} p_0 q_0$	1950 $\frac{p_n}{p_0} p_0 q_0$	1952 $\frac{p_n}{p_0} p_0 q_0$
Creamery butter	100.0	101.1	118.7	72.39	72.39*	73.19	85.93
American cheese	100.0	101.7	126.7	27.11	27.11	27.57	34.35
Condensed milk	100.0	100.9	117.8	77.03	77.03	77.72	90.74
Fluid milk	100.0	96.0	114.7	460.77	460.77	442.34	528.50
Eggs	100.0	84.0	93.6	195.00	195.00	163.80	182.52
Total				832.30	832.30	784.62	922.04
				$\Sigma p_0 q_0$	$\Sigma \frac{p_0}{p_0} p_0 q_0$	$\Sigma \frac{p_n}{p_0} p_0 q_0$	$\Sigma \frac{p_n}{p_0} p_0 q_0$
$\frac{\Sigma \frac{p_n}{p_0} p_0 q_0}{\Sigma p_0 q_0}$						784.62	922.04
						832.30	832.30
Index					100.0	94.3	110.8

* Since a price relative is in percent, in computation it must be used in decimal form. Since 100% = 1.00, we have here $72.39 \times 1.00 = 72.39$, and so on.

Source: Table 18.2 and Table 18.3.

If the transformed formula is applicable, the results are the same for the weighted average of price relatives and for the weighted aggregate of actual prices, but the procedures are different. Therefore, let us illustrate the procedure of using the untransformed formula, as in Table 18.4 where we have taken the price relatives from Table 18.2 and the value weights from Table 18.3. If these were not already available the five following steps would have to be taken:

1. Compute price relatives, p_n/p_o .
2. Compute value weights, in this instance using $p_o q_o$.
3. Multiply each relative by the appropriate weight, that is, obtain $\frac{p_n}{p_o} p_o q_o$.

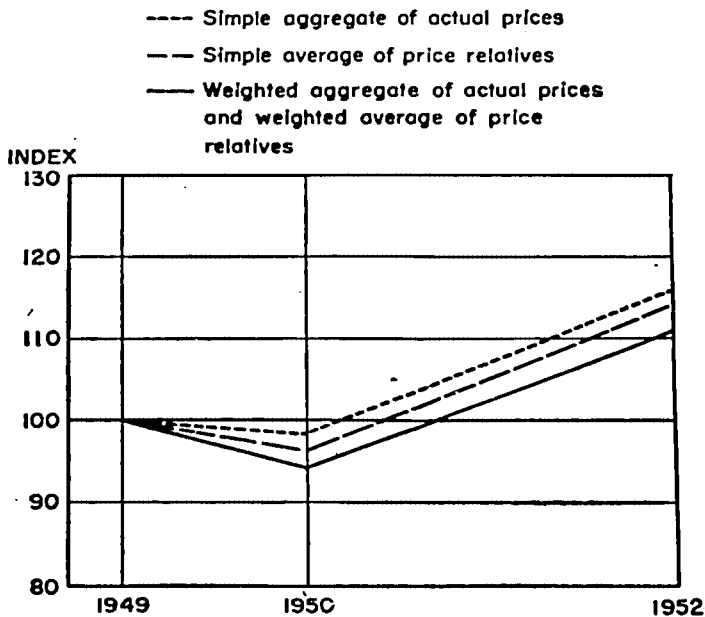


Chart 18.1. Wholesale Price Index of Dairy Products in the United States for 1949, 1950, and 1952 (Comparison of Results of Various Methods). 1949 = 100.

Source: Tables 18.1, 18.2, 18.3, and 18.4.

4. Total the products: $\Sigma \left(\frac{p_n}{p_o} p_o q_o \right)$
5. Divide these totals by the sum of the weights, $\Sigma p_o q_o$.

This procedure is symbolized by the untransformed formula, which is here repeated:

$$P = \frac{\Sigma \left(\frac{p_n}{p_o} p_o q_o \right)}{\Sigma p_o q_o}$$

In Chart 18.1 are shown the indexes obtained by the four methods we have discussed. There is marked difference in the results obtained by the several methods. The weighted methods are superior to the simple methods for reasons already discussed.

SPECIAL PROBLEMS

Among the special problems that arise in the use and construction of index numbers are : (1) making differently based index numbers comparable; (2) combining index numbers; (3) splicing series of index numbers; and (4) expressing percentage changes in index numbers. These four special problems will be briefly explored.

Making Indexes Comparable

In order to compare, for example, the "cost-of-living" index of the Bureau of Labor Statistics and that of the National Industrial Conference Board for 1953 and 1954, it is necessary to establish them on the same base. This problem is actually one of shifting the base. The BLS index is on a 1947-1949 base; the NICB index is on a 1953 base. We can either recompute the BLS index for 1953 and 1954 in terms of its 1953 base, or recompute the NICB index for 1953 and 1954 in terms of its 1947-1949 base. To illustrate, let us do the latter. Then the average

value of the NICB index for 1947-1949 must be found, and this average for 1947-1949 becomes the new base for the NICB index. To transform the NICB index numbers on the 1953 base to the new 1947-1949 base we must divide them by the new base.

The NICB index numbers for 1947-1949 are totaled and divided by the number of years. These index numbers, as shown in Table 18.5, are as follows: 84.7, 90.1, 88.8. They total 263.6,

TABLE 18.5. COMPARISON FOR 1953 AND 1954 OF THE BUREAU OF LABOR STATISTICS CONSUMER PRICE INDEX AND THE COST-OF-LIVING INDEX OF THE NATIONAL INDUSTRIAL CONFERENCE BOARD.

Year	BLS index (1947-1949 = 100)	NICB index (1953 = 100)	NICB index (1947-1949 = 100)
1947		84.7	
1948		90.1	
1949		88.8	
1950		90.0	
1951		97.0	
1952		99.5	
1953	114.4	100.0	113.8
1954	114.8	100.2	114.0

Source: *Survey of Current Business*, 1953 supplement and issue of February 1955; *National Industrial Conference Board Management Record*, May 1955.

which being divided by 3 gives 87.9. We divide the NICB index 100.0 for 1953 by 87.9. This gives us 1.138 or 113.8%; that is, 113.8 is the 1953 NICB index if 1947-1949 = 100.* For 1954 this procedure gives us 114.0.

Thus, we have made 87.9 (the average of the NICB's index numbers for 1947 through 1949) the new base, and expressed the 1953 and 1954 index numbers as percents of 87.9. This is the same as considering 87.9 as 100 and comparing 1953 and 1954 figures to it.

* Index numbers in any computation are used in their decimal form rather than in their percentage form. Thus an index of 83 is for computational purposes .83.

The BLS figures for 1953 and 1954 on the 1947-49 base are 114.4 and 114.8 respectively. These figures are very close to the NICB figures when expressed on their 1947-1949 base.

Combining Index Numbers

Let us assume that we have an index of weekly employee earnings for the Suprema Corporation, which shows an increase from 1950 to 1952. In addition, we have an index of the "cost of living" during the same time span. In Table 18.6, these two indexes are shown. Our aim is to combine this information into an index of real wages. This actually is a process of "deflating." We divide the index of weekly employee earnings for each year by the Consumer Price Index for that year, and achieve an index of real wages, as shown in Table 18.6. Thus, we are really adjusting nominal wages for changes in the consumers' prices by dividing.

TABLE 18.6. ADJUSTMENT OF NOMINAL WAGES IN THE SUPREMA CORPORATION OF CANTON, OHIO, FOR CHANGES IN THE BLS CONSUMER PRICE INDEX, 1950-1952 (1947-1949 = 100).

<i>Year</i>	<i>Index of Employees' Weekly Earnings, Suprema Corporation</i>	<i>BLS Consumer Price Index</i>	<i>Index of Real Wages in the Suprema Corporation</i>
1950	105.6	102.8	102.7
1951	108.6	111.1	97.7
1952	110.3	113.5	97.2

The result in our illustration is that real wages do not show the increase that nominal wages show.

Splicing

If new commodities have entered the market, we may wish to construct an index number including these commodities, but continuous with the old index. To accomplish this con-

tinuity requires that we "splice" the indexes. The procedure is illustrated in this example.

<i>Year</i>	<i>First Index Number</i>	<i>Second Index Number</i>	<i>Spliced Index Number</i>
1952	100		100
1953	115		115
1954	122		122
1955	130	100	130
1956		110	143

We splice at the year where the new index has its base. In our illustration we establish a relationship for the year at which we splice—here 1955—between the first index number, which is 130, and the second index number, which is of course 100. This relationship then prevails for the years that follow, as for example, 1956.

$$1.00 : 1.30 = 1.10 : x \quad \text{and} \quad x = 1.43.$$

Thus, in our illustration, all future values of the second index number are multiplied by 1.3 if we want to achieve an index continuous with the first index.

Percentage Change in Index Numbers

It is important to state what percent of change has occurred in price index numbers from one time period to another. The same change may be expressed in two different percentage figures, and both may be correct. But it is necessary to make clear which of the two methods has been used.

Thus, the United States Wholesale Price Index for 1944 is 67.6, and for 1954 is 110.3. We wish to express the change that has occurred in the United States Wholesale Price Index from 1944 to 1954. We may express it as a difference or as a ratio, that is

$$110.3 - 67.6 = 42.7\% \text{ change}$$

or

$$\frac{110.3 - 67.6}{67.6} = \frac{42.7}{67.6} = 63.2\% \text{ change}$$

In the first equation, the change is merely the difference in percentage points. In the second equation, the change is a ratio of the difference between the 1954 figure and the 1944 figure, to the 1944 figure, expressed as a percent.

QUANTITY INDEX NUMBERS

Price index numbers measure, and permit comparison of, the price of certain goods; quantity index numbers, on the other hand, measure, and permit comparison of, the physical volume of goods produced, or distributed, or consumed. The index of quantity produced is the leading type of quantity index number. Though price indexes are more widely used, production indexes are highly significant as indicators of the level of output in the economy or in parts of it.

The direction in which the economy is moving is shown by how much is being produced and how this level of production compares with previous levels. One reads in newspapers such statements as the following from the United Press on July 3, 1950: "Unlike 1941, there is little room for added military production without cutting into consumer goods production. . . . In May the Federal Reserve Board's [production] index stood at 193 and was expected to reach or surpass [in June] the previous peacetime record of 195."

In measuring the standard of living in terms of physical output, quantity indexes are invaluable.

In constructing quantity index numbers, the problems confronting the statistician are analogous to those involved in price indexes. We measure changes in quantities, and when we weight we use prices or values as weights.

The *simple aggregate of quantities* has the formula

$$Q = \frac{\sum q_n}{\sum q_o}$$

This type of quantity index cannot always be used since it requires that all data be in the same units. Obviously, we cannot add tons, pounds, and gallons.

The *simple average of quantity relatives*, when the arithmetic mean is used, is found through the formula

$$Q = \frac{\sum \left(\frac{q_n}{q_o} \right)}{N}$$

The *base-year weighted aggregate of quantities* bears the formula

$$Q = \frac{\sum q_n p_o}{\sum q_o p_o}$$

The *base-year weighted average of quantity relatives* (when the arithmetic mean is used) has the formula

$$Q = \frac{\sum \left(\frac{q_n}{q_o} p_o q_o \right)}{\sum p_o q_o}$$

The limitations and conditions for interpretation that apply to the different kinds of price index numbers apply analogously to quantity index numbers.

VALUE INDEXES

The value index is a type not in wide use. The index of department store sales of the Federal Reserve Board is one of the most important value index numbers currently being compiled.

Value means price times quantity. Thus a value index V is the sum of the values of a given year divided by the sum of the values of the base year. The formula therefore is:

$$V = \frac{\sum p_n q_n}{\sum p_o q_o}$$

Since in most cases the value figures are given, the formula may be stated more simply

$$V = \frac{\sum v_n}{\sum v_o}$$

in which v stands for value.

In this type of index both price and quantity are variable in the numerator. Weights do not have to be applied since they are inherent in the value figures. A value index therefore is an aggregate of values.

SPECIAL-PURPOSE INDEXES

Under this heading fall a large variety of indexes. The *New York Times* Index of Business Activity is one illustration. In special-purpose index numbers various kinds of data are often combined. There is a great variety of highly specialized indexes and the methods of constructing them are very varied also. In the next chapter illustrations of special-purpose index numbers will be given.

Summary

1. Index numbers are widely used statistical devices.
2. There are single-series index numbers and composite index numbers.
3. The concepts of ratio, average, dispersion, sampling, and time series are involved in index numbers.
4. Index numbers may relate to price, to quantity, to value, or to a special purpose.
5. The basic problems confronting the statistician in constructing index numbers concern data, base, combining of the data, and weighting of the data.
6. Four important methods of constructing price index numbers are the simple aggregate of actual prices, the simple average of price relatives, the weighted aggregate of actual prices, the weighted average of price relatives.

7. Some special problems are (a) to make differently based index numbers comparable; (b) to combine index numbers; (c) to splice indexes; (d) to express percentage changes in index numbers.

8. Quantity index numbers are constructed analogously to price index numbers.

9. Value indexes and special-purpose indexes are distinct types which have practical significance.

CHAPTER

19

Current Indexes

IMPORTANT PRICE INDEXES

The two most important price indexes in the United States today are the Wholesale Price Index and the Consumer Price Index, both compiled by the same government agency, the Bureau of Labor Statistics.

The following descriptions are based upon publications of the Bureau of Labor Statistics, and in many instances consist of unchanged statements of the Bureau.*

Wholesale Price Index

In Chart 19.1 the Bureau of Labor Statistics index of wholesale (primary market) prices is shown for the period 1939 to 1954.

How is this statistical measuring rod constructed and used?

• **HISTORY OF THE INDEX.** Since 1902, the official monthly † wholesale price index has been prepared as an indicator of general price trends and average changes in commodity prices at primary market levels. The official series has been carried back to 1890; separate monthly indexes are available for major groups of

* "A Description of the Revised Wholesale Price Index." *Monthly Labor Review*, February 1952; *The Consumer Price Index, A Short Description of the Index as Revised*, 1953, United States Department of Labor; *The Consumer Price Index*, Bulletin No. 1140, United States Department of Labor, 1953.

† In recent years, a weekly wholesale price index and a daily index of "spot market" prices have been issued.

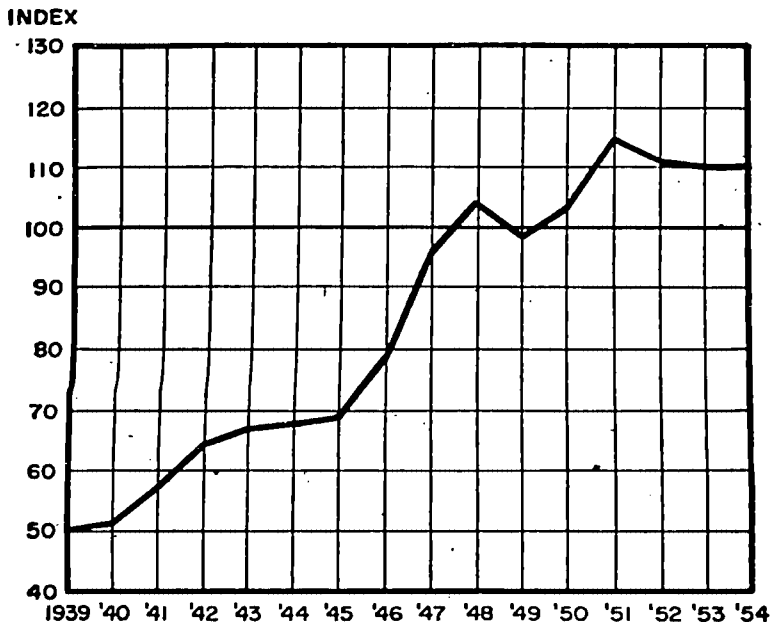


Chart 19.1. United States Wholesale Price Index, 1939–1954
(1947–1949 = 100).

Source: *Monthly Labor Review*.

commodities from January 1890, and for subgroups of commodities beginning with 1913. The composite index has been estimated, jointly by the Bureau and by nongovernmental agencies and individuals, on an annual basis back to 1749.

During recent decades, theoretical work in the field of index-number construction, together with empirical studies of price relationships, has pointed toward the need for and the advantages of a more precise definition and statement of principles, as well as certain refinements in the index structure. A revised index of wholesale prices extending back to January 1947 was issued by the Bureau of Labor Statistics early in 1952. The review and modernization work on this index dealt with all aspects, including the classification system, base period, weights, and calculation method.

CONTENTS AND USES OF THE INDEX. "Wholesale" as used in the title of this index refers to sales in large lots, not to the prices paid or received by wholesalers, distributors, or jobbers. The price data used in constructing this index are those which apply at primary market levels—that is, the level of the first important commercial transaction for each commodity in the sequence from producer to consumer. Most of the quotations are the selling prices of representative manufacturers or producers, or prices quoted on organized exchanges or markets.

The index is designed to show the general rate and direction of the composite of price movements in primary markets and the specific rates and directions for individual commodities or groups of commodities. It is intended to measure "real" or "pure" price changes between two periods of time, that is, to measure price changes not influenced by changes in quality, quantity, terms of sale, etc.

Currently the index contains prices of approximately 2000 commodities classified by product, rather than by industry or end use, into 15 major groups and 88 subgroups. Each subgroup is broken down into "product classes."

Each group or subgroup consists of a related group of products alike as to raw material, end use, or production process. However, the products within each subgroup are not necessarily of the same economic class—some subgroups consist entirely of highly fabricated products such as construction machinery; others, like cotton products, range from semifinished goods (yarns) through to final finished products (sheets or curtains). A product class approximates a grouping of commodities, produced by one or more related industries, and which are also characterized by similarity of price movement, raw materials, or production process.

The fifteen major groups are:

1. Farm products.
2. Processed foods.
3. Textile products and apparel.

4. Hides, skins, and leather products.
5. Fuel, power, and lighting materials.
6. Chemicals and allied products.
7. Rubber and products.
8. Lumber and wood products.
9. Pulp, paper, and allied products.
10. Metals and metal products.
11. Machinery and motive products.
12. Furniture and other household durables.
13. Nonmetallic minerals—structural.
14. Tobacco manufactures and bottled beverages.
15. Miscellaneous.

The major uses of the index, of its components, and of the individual price series are numerous. They provide one tool in the analysis of general business conditions and of the business cycle. In conjunction with other data, they indicate the state of balance between aggregate supply and demand at the primary market level. Individual concerns utilize the indexes in planning plant construction and production schedules, in purchasing raw materials, in evaluating inventories, in determining production costs, and in general investment programs. The indexes are widely used as bases for adjusting prices in contracts extending over a period of time—"contract escalation." Other uses are for forecasts of future pricing, marketing, and sales policies based on historical price developments; for analysis of the price structure of the economy and changing interrelationships among individual commodities; and in combination with other data, for formulation of long-range economic and other programs by government officials.

METHOD OF COLLECTION. The index is based upon a sample* of commodities, specifications, reporters, and primary-market levels of transaction (see below). The commodities included in the index are not random samples. The selections are based upon knowledge of each industry and its important products,

* See Chapters 21-23 for an extended discussion of samples and sampling.

and are usually made after consultation with leading trade associations and manufacturers within each field. In general, the commodities included in the index are the most important ones in each field (using Census and similar data to determine importance); some, not important in terms of sales volume, appear to offer good representation of price movements because of certain industry or trade characteristics.

Detailed specifications have been written for the items priced. Each specification has been prepared in cooperation with experts from industry and trade, to be as typical as possible of the price movements of all specifications for the commodity. An example of specifications is:

Alcohol, butyl, normal, synthetic, tanks, f.o.b. works, freight allowed East, per pound.

In selecting the price reporters, informed judgment and industry advice are relied upon. If an organized exchange or market represents the industry or commodity, the exchange or market price is used. If a trade journal quotation is generally recognized as reliable, it is used. In other instances, as for most agricultural products and fish, prices for goods of stated specifications are officially collected and published by other government agencies. For quotations from individual producers (the Bureau does not use prices from buyers) reporters are selected, taking into account the distribution of companies by geographic region (where such distribution is important), by degree of integration, by volume, and by degree of "price leadership." Whenever possible, at least three reporters are obtained for each specification.

CALCULATION PROCEDURES. The Wholesale Price Index is calculated by adaptation of the Laspeyres formula (see page 315 herein). The base period for prices is an average of the three years 1947, 1948, 1949, but for the vast majority of commodities weights are based on transactions in 1947. In order to keep the weights from becoming outmoded, the Bureau of Labor Statistics periodically reviews the weights in the light of data which become available.

LIMITATIONS. The index measures completely the price changes of only the approximately 2000 commodities which are specifically defined for pricing. It is not an over-all measure of the "general price level" or of "the purchasing power of the dollar"—it does not include all classes of commodities which are factors in the "general price level." The index does not measure the price movements of retail transactions, transactions for services (including construction of finished structures or of other composite items which are not reproduced on a repetitive base), printing and publishing, real estate, transportation, or securities. For certain commodities, such as ships and railroad rolling stock, fabricated plastic products, and some machinery which is largely custom-made, it has not been possible for the Bureau of Labor Statistics to obtain any reliable measures of price movement.

Consumer Price Index

In Chart 19.2 the United States Consumer Price Index of the Bureau of Labor Statistics is shown for the years 1939 to 1954.

HISTORY OF THE INDEX. The index was initiated during World War I, when prices rose rapidly, for use in wage negotiations, particularly in shipbuilding centers. In the fall of 1935, the Bureau introduced improved methods of calculating the index and in 1940 completed revision of the weights to correspond with 1934-1936 family expenditure patterns, as determined by extensive study of family consumption. In 1953, after extensive research, a revised index was issued.

WHAT THE INDEX MEASURES. The Consumer Price Index of the Bureau of Labor Statistics is a statistical measure of changes in prices of the goods and services bought by families of city wage earners and clerical workers. It measures only *changes in prices*; it tells nothing about changes in the kind and amounts of goods and services families buy, or the total amount families spend for living, or the differences in living costs in different places. Other indicators are required to measure these facts.

The complete name for this index is: *Index of Change in Prices*

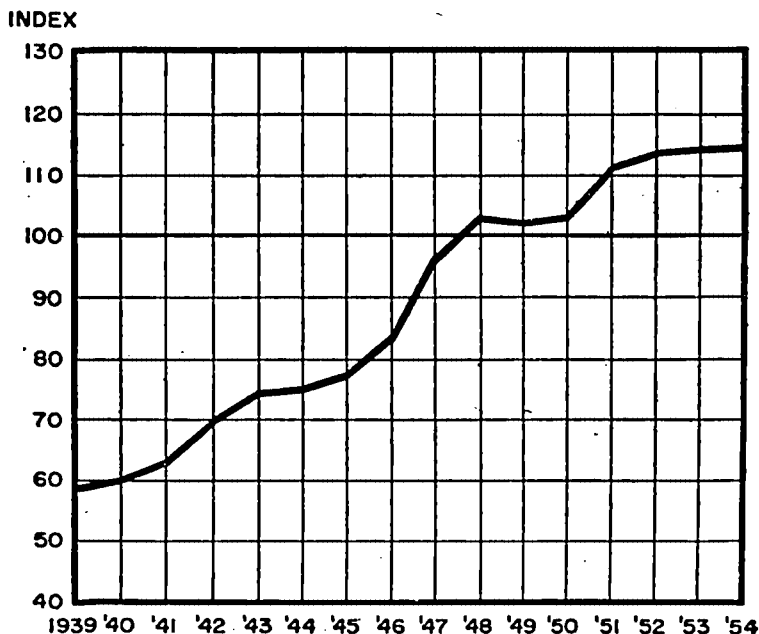


Chart 19.2. United States Consumer Price Index, 1939-1954 (1947-1949 = 100).

Source: *Monthly Labor Review*.

of Goods and Services Purchased by City Wage-Earner and Clerical-Worker Families to Maintain Their Level of Living. This long title contains the key to the meaning of the index.

The index is concerned with *prices*. These are the retail prices of foods, clothing, housefurnishings, fuel, and other goods; the fees paid to doctors and dentists; prices in barber shops and other service establishments; rents; rates charged for transportation, electricity, gas, and other utilities; and so on. They are prices paid by consumers, and they include sales and excise taxes.

The "market basket" of goods and services priced for the index is representative of the goods and services bought by urban wage-earner and clerical-worker families. These are families of two or more persons living in the 3000 towns, cities, and suburbs of the United States, ranging in population from

small cities of 2500 people to New York City. The heads of these families are wage earners or salaried clerical workers, including craftsmen, factory workers, laborers, clerks, sales and service workers (except domestic-service workers). Many of the families have two or more wage or salary workers; as a result, average *family incomes* are higher than average *individual earnings*. But families with incomes of \$10,000 after taxes are excluded.

Thus, this market basket is an estimate of the goods and services bought by this group of the population in 1952 to use, replace and add to their possessions, for the purpose of keeping up their 1952 level of living. Wage-earner and clerical-worker families are "moderate income" families; thus, the market basket does not reflect purchases of families at the lowest and highest extremes of the income scale, who have different scales of living.

The index is concerned with *change* in prices. It measures the effect of price changes on the cost of the goods and services in the market basket. The content of the market basket—that is, the quantities and qualities of goods and services that represent what families bought in 1952—is assumed to remain essentially the same, so that the change in cost from month to month is the result of *changes in prices*. If families buy poorer or better things, or buy more or less than they bought in 1952, the index does *not* undertake to measure the change in their spending that results from these changes in their level of living; it measures *only* the change in spending caused by *changes in prices*.

USES OF THE INDEX. Extensive use of the Consumer Price Index led a congressional committee to term it, in 1951, the government's "most important single statistic." Its most widely publicized use is in the automatic adjustment of wages under "escalator clauses" in collective-bargaining agreements. Its use in collective bargaining is actually more widespread. Even where automatic escalator clauses are not in effect, the index is a factor in determining the wage rates of many workers. The index is also used, both directly and indirectly, by Federal, State, and local government agencies in setting pay scales for their employees. In

1952 it was estimated that wages of several million employees were adjusted according to changes in the Consumer Price Index.

In addition, both the Congress and the executive agencies of the Federal Government make wide use of the index as a guide in determining general economic policy, particularly with respect to wages, prices, and taxation.

BASIS FOR SELECTION OF ITEMS. To determine what goods and services wage-earner and clerical-worker families buy, the Bureau made a Survey of Consumer Expenditures in 1950. The information thus obtained from about 8000 wage earners' and clerical workers' families was analyzed and averaged to determine: (1) the kinds of goods and services bought; (2) the quantities bought and the amount spent for each article; and (3) the quality of each article bought. These averages were then adjusted for changes in prices and buying habits that occurred between 1950 and 1952.

The market basket is based on the 1952 figures and contains the following broad classes of goods and services:

Food	Medical care
Housing	Personal care
Apparel	Reading and recreation
Transportation	Other goods and services

PRICES AND SPECIFICATIONS. In 1950 and 1951 the Bureau priced and studied the price changes of hundreds of items. Using this information and extensive price records of the past, the Bureau selected about 300 items, which together can be used to estimate the average change in prices of all items in the market basket. Among these items are all the goods and services that are of outstanding importance in family purchases, so that the priced items in themselves represent the greater part of family spending. Prices of these items are collected at regular intervals and the successive prices are compared to calculate price changes. It is important to be certain, however, that the calculation shows *only* price change and not changes due to such other things as quantity or quality differences. For example, the price

of pork chops can vary considerably depending on how near the center of the loin they are cut. If center-cut pork chops are priced one month and end-cut pork chops the next month, the price of pork chops might *seem* to have gone down. But in fact this price difference would reflect a difference in *quality* and not a real change in *price*. To prevent this confusion insofar as is possible, the Bureau has prepared detailed specifications to describe the items that are priced for the index. These specifications are carefully written, with the advice of manufacturers and retailers who know the articles. For example, the following is a specification for one of the men's shirts priced:

<i>Style:</i>	business; fused or similarly constructed collar, attached; barrel cuffs;
<i>Fabric:</i>	cotton broadcloth, white;
<i>Yarn:</i>	combed;
<i>Thread count:</i>	136 × 60 or 128 × 68;
<i>Finish:</i>	residual shrinkage 1 percent or less;
<i>Construction:</i>	full cut; fair workmanship; 31 to 32 yards per dozen based on 36-inch fabric;
<i>Size range:</i>	14- to 17-inch neckband;
<i>Brand:</i>	manufacturer's nationally advertised.

METHODS OF COLLECTION. From 97 cities covered by the expenditures surveys, the Bureau selected a representative sample of 46 cities in which prices are collected: the 12 largest, 9 other large cities, 9 medium-sized cities, and 16 small cities. In each city, the Bureau selected a list of stores and all other types of establishments where families of wage and salary workers buy goods and services. This list includes representative chain stores, independent stores, department stores, and specialty stores. Prices reported by these stores are averaged together for each city to determine average price changes. Where people do more buying in independent stores, independent-store prices are given more importance in the average; where they do more buying in chain stores, chain-store prices are given more importance. Periodic checks are made to see that the sample of stores in each city is kept representative.

Prices are collected in each city at intervals ranging from once every month to once every four months. Because food prices change very frequently and because foods are a very important part of total family spending, food prices are collected in *each* city *every* month. A few other important items are also priced each month in each city; for example, cigarettes, fuels, and gasoline. Prices of other goods and services are collected every month in the five largest cities. In other large cities and in the medium-sized cities, prices for items other than foods, fuels, and rent are obtained once every three months; and in small cities they are collected once every four months.

Food prices are collected around the middle of each month by local part-time agents, often housewives, who are trained to follow standard pricing procedures and specifications. Commodities and services other than food are priced by the Bureau's full-time field agents.

Rent information is collected by mail every month from each city. Then every year or two the Bureau's agents visit the homes of a large representative sample of renting families to verify mail reports on rent and facilities such as heat, electricity, etc., included in rent payments. During the personal collection of rent information, the Bureau brings its sample of renters up-to-date to include new dwellings.

In addition to rents, a few other prices can be collected successfully by mail. The Bureau uses mail questionnaires to obtain street-car and bus fares, public-utility rates, newspaper prices, prices of fuels and certain items, which change in price only occasionally and do not require personal visits by Bureau agents.

CALCULATION PROCEDURES. The current base period is 1947-1949. The weights are 1951-1952 quantities. The index is basically a weighted aggregate of actual prices, using an average of weights of typical years. But in practice the Bureau of Labor Statistics calculates the index on a variation of this formula as a weighted average of price relatives for each item, the weights being the value weights of the typical years.

LIMITATIONS. The Consumer Price Index is necessarily an

approximation. It is based on a sample of families, items, stores, and cities that are a fraction of the total. Even time is sampled in the sense that information is collected only at certain periods. Thus sampling errors are inevitable. Since respondents do not usually recall exactly all of their expenses, the index is subject to response errors. Despite the careful training of price agents of the Bureau of Labor Statistics, there are situations like quality changes, special discounts, and premium prices which occasionally escape their attention.

The above limitations inhere in its construction. There are also limitations in the applicability of the Consumer Price Index. The Consumer Price Index is calculated specifically to measure the average change in prices of goods and services bought by urban wage-earner and clerical-worker families. Consequently, it must be used carefully when it is used for other purposes. The index represents *all* wage-earner and clerical-worker families, but not necessarily any *one* family or any *small group* of families. There are limitations on the application of the index to very low or very high income groups, to elderly couples, to single workers, or to other groups whose level or manner of living and spending are different from the average of all worker families. To the extent that these groups spend differently and are therefore differently affected by price changes, the index is not exactly applicable. On the other hand, when the index is applied to *all* city families or to the *total* urban population, the limitation may not be serious, since the wage-earner and clerical-worker family group represents such a large proportion (nearly two-thirds) of these populations.

Neither should the index be used to measure the changes in *total* family *spending*, since it measures only the effect of *price* change and does not take into account other factors, such as higher or lower incomes or income taxes. The index does not reflect the experience of the individual housewife as she "shops around" to take advantage of the lowest prices, nor does it show the full effect of paying premium prices for scarce items. And the index does not reflect the change in costs experienced

by families who move from one city to another or who change from renting to owning their own homes.

Other Price Indexes

Other price indexes are currently being compiled and used. Some are compiled by government agencies and some by private organizations. Among the latter are the Wholesale Food Price Index of Dun and Bradstreet. This index uses the method of simple aggregate of actual prices expressed in dollars rather than percent. Another is the Retail Price Index for Textiles and Textile Products compiled by Fairchild Publications of New York City. A nongovernmental cost-of-living index is that of the National Industrial Conference Board.

IMPORTANT QUANTITY INDEXES

The Federal Reserve Board's Index of Industrial Production *

In Chart 19.3 the index of Industrial Production of the Federal Reserve Board is shown from 1939 to 1954.

CONTENT AND PURPOSE OF THE INDEX. The index of industrial production has been designed to measure monthly changes in the physical volume of output of manufacturing and mining industries. These industries, covering a vast and heterogeneous array of products and processes, account for approximately one-third of the national income. The index of industrial production not only measures directly the output of industries accounting for an important part of the variations in total economic activity, but also, since these industries represent strategic sectors of the economy, it quickly reflects changes likely

* *Federal Reserve Monthly Index of Industrial Production*, 1953 Revision, *Federal Reserve Bulletin* for December 1953, issued by the Board of Governors of the Federal Reserve System, Washington, D. C., 1953; and "New Federal Reserve Index of Industrial Production." *Monthly Review of Credit and Business Conditions*, Federal Reserve Bank of New York, January 1954, volume 36, No. 1. In many instances we have taken over, unchanged, the statements of the Board.

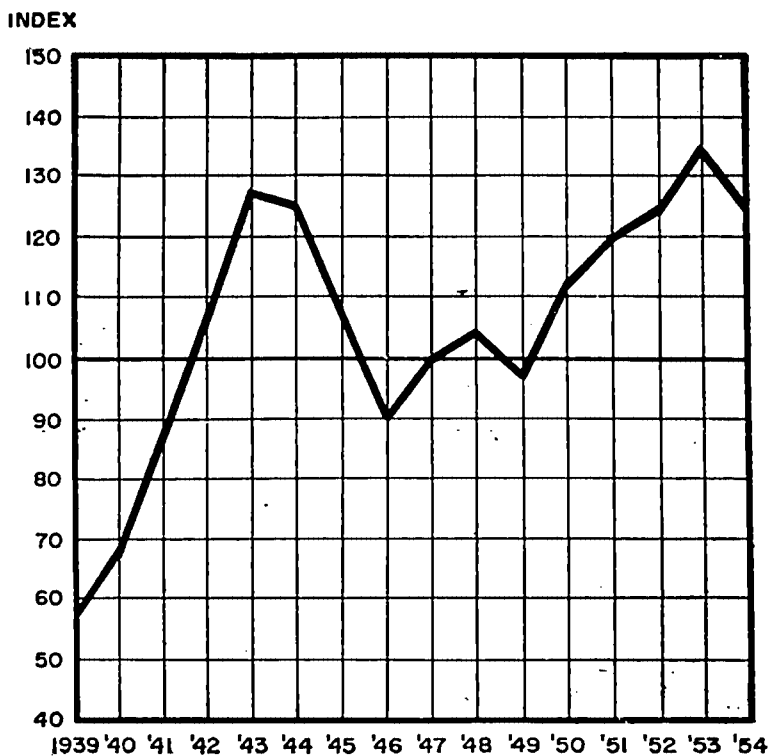


Chart 19.3. Index of Industrial Production in the United States, 1939-1954 (1947-1949 = 100).

Source: *Federal Reserve Bulletin*.

to have a wide-reaching influence upon over-all economic activity.

Two forms of the index of industrial production are now compiled, one a monthly index based upon 175 series and the other an independent annual index based on approximately 1400 series. The average relative importance of the major industry classifications in the monthly series are: durable manufactures, 45.17%; nondurable manufactures, 44.85%; and minerals, 9.98%. Earlier versions of the index tended to reflect more heavily the activity of basic industries, which fluctuate more sharply than nonbasic industries. The current version of the

index makes use of more data representing intermediate and final stages of production.

The monthly index is published with and without seasonal adjustment. The monthly and annual indexes are broken down into groups and subgroups which make it possible to study and compare changes in different sectors of the economy.

CALCULATION AND COLLECTION PROCEDURES. The method used in compiling the Federal Reserve Board's Index of Industrial Production is that of the weighted average of quantity relatives. The base period is the average for the three years 1947-1949. The use of 1947-1949 as a base does not imply that that period was a "normal" one with regard to industrial production. It is the period recommended by the Office of Statistical Standards of the Bureau of the Budget and is now used by many statistical agencies, government and private. The adoption of this base period facilitates comparison with many other indexes and makes this index convenient to work with.

Weights for the index are based largely on value-added data. These value-added weights give the value added by manufacture to the raw materials used and are derived, in the main, from the 1947 Census of Manufactures.

In computing the index the Federal Reserve Board makes use of a wide variety of statistical series compiled by both government and private sources. Most of the data are gathered by United States government agencies.

Since the index is intended to measure physical output, the preferred type of data is that which measures quantities produced in physical terms. The index is not rigidly limited to this type of measurement of output, however, mainly because for a great many industries the only types of data available on a monthly basis do not measure physical output. Industrial production in certain industries, shipbuilding for example, cannot be measured by monthly output since in them the time period for completed production is relatively long. Where it is not possible or advisable to use physical units produced as the

measure of output, related series which tend to fluctuate more or less closely with output are used, such as shipments, materials consumed, and man-hours. The annual index relies for the most part upon measures more closely related to physical production while the monthly index makes relatively greater use of man-hour data.

LIMITATIONS. Among the more important areas of the economy not covered by the index are construction activity, public utilities, transportation, trade, services, and agriculture. The index is not, therefore, a measure of total economic activity.

Wherever production data are estimated not through physical output but through some related series such as man-hours, inaccuracies may result.

Even with the improved procedures initiated in the revision of 1953, the index cannot allow immediately for current changes in seasonal patterns, nor make prompt and accurate adjustments for the emergence of seasonals in industries which have not experienced such variation through years of sustained peak production.

World Index of Industrial Production

The Statistical Office of the United Nations publishes a World Index of Industrial Production relating to mining and manufacturing. Construction and the production of gas and electricity are not included in this index. Separate regional indexes are also published along with this world index. The World Index of Industrial Production does not include data on production in the U.S.S.R., Bulgaria, China, Czechoslovakia, Hungary, Poland, Romania, and East Germany.

The index numbers in the United Nations World Index of Industrial Production have been computed as the arithmetic means of national indexes on the base 1948 = 100 weighted according to value added in 1948 in the mining and manufacturing industries. Recently the base was shifted to 1953 = 100 by dividing all index numbers by 137, which was the index number for 1953 when 1948 was the base. For the years 1946-1955 the

World Index of Industrial Production of the United Nations was as follows:

	1948 = 100	1953 = 100
1946	83	61
1947	93	68
1948	100	73
1949	101	74
1950	115	84
1951	125	91
1952	128	93
1953	137	100
1954	—	100
1955	—	110

Further details concerning the World Index of Industrial Production can be found in issues of the *Statistical Yearbook of the United Nations*.

VALUE INDEXES

A good example of a value index is the Index of Department Store Sales prepared by the Federal Reserve Board.*

A sample of department stores from each of the twelve Federal Reserve Districts report their sales each month. Reports are obtained from virtually every important department-store trading area in the country. The sample of the sales index includes approximately 1300 stores. The dollar-volume coverage of the sample is currently 84% of total sales of all department stores as presently defined. The samples of reporting stores are somewhat larger than would be required to measure national trends in department-store sales, since the Federal Reserve Board has found it desirable to have information not only on national business developments but on district and local business developments as well; and, in order to provide information on this

* "Revised Indexes of Department Store Sales and Stocks," *Federal Reserve Bulletin*, December, 1951.

basis, a much broader sample is required. The information gathered for the index has been found to be of considerable use to the Federal Reserve System and also to the respondent stores, trade associations, research organizations, and others concerned with regional developments.

Indexes of sales are first computed for each of the twelve Federal Reserve Districts and then combined into a national index. Before combining the district indexes, each of them is weighted according to the relationship of total department-store sales in each district to the total for the country as a whole in the base period. The base period is 1947-1949.

The index of department-store sales measures the level of average daily sales. The monthly sales of the stores included in the index sample are summed and these monthly totals are then divided by the number of trading days in the month to obtain average daily sales. The use of average daily sales largely eliminates the fluctuations in the volume of sales resulting from calendar irregularities.

The average daily sales are converted into index numbers by dividing each by the average daily sales during the base period, 1947-1949. The district indexes are seasonally adjusted before combining them into a national index. A special adjustment for the changing date of Easter is also necessary.

Special-Purpose Indexes

Under this heading falls a large variety of indexes. The *New York Times* Index of Business Activity has been mentioned. Also, there are various trade and industry barometers. Some special-purpose indexes measure commercial or industrial efficiency; for example, there is the Index of Railroad Operating Efficiency computed by the Bureau of Railway Economics of the Association of American Railroads. In addition, there are indexes of sales opportunities, usually by geographical areas, computed by market-research organizations.

Other special-purpose indexes that should be mentioned are the labor indexes of the Bureau of Labor Statistics. An example

of these is the Index of Production-Worker Employment in Manufacturing Industries.

Summary

1. Two important price indexes are the United States Wholesale Price Index and the United States Consumer Price Index, both compiled by the Bureau of Labor Statistics of the United States Department of Labor.

2. Important price indexes also are compiled by nongovernmental agencies.

3. The best-known quantity index is the Index of Industrial Production of the Federal Reserve Board. A World Index of Industrial Production is published by the United Nations.

4. Value indexes are well exemplified by the index of Department Store Sales of the Federal Reserve Board.

5. In addition to the indexes mentioned, there is a large number of what are called special-purpose indexes.

Introduction to Correlation

The Concept of Correlation

The board of directors of the Excelsior Corporation, operating the chain of Excelsior Department Stores, is faced with the problem of estimating what the annual sales might be in a store to be opened in Mill City, where Excelsior has not heretofore operated. They need this information in order to plan the size of store, the amount of stock to be put on the shelves, the number of employees to be hired. An answer to this problem can be found statistically by establishing the general relationship, in the cities in which the corporation is already operating, between the size of a city's employed labor force and sales in its Excelsior stores. From the size of Mill City's employed labor force, which is 100,000, we can estimate what the annual sales will be in the new store, and the board of directors should be able to base its decision on this statistical estimate. But the board will also want to know how good this estimate is, since it is based only on a general relationship between sales and employment. Still further, the board may want to compare the relationship existing between sales and labor force on the one hand with the relationship existing between sales and number of stores on the

other hand. Such problems can be solved through correlation analysis. The raw data are found in Table 20.1.

TABLE 20.1. 1950 SALES OF THE EXCELSIOR DEPARTMENT STORES AND APPROXIMATE EMPLOYMENT FIGURES IN 12 CITIES.

<i>City</i>	<i>Employment (Thousands of Workers)</i>	<i>Annual Sales (Thousands of Dollars)</i>
Abilene, Tex.....	22	250
Alexandria, La.....	31	200
Charleston, W. Va.....	90	980
Evansville, Ind.....	82	850
Fort Smith, Ark.....	43	710
Jackson, Miss.....	65	280
Roanoke, Va.....	59	630
Sante Fe, N. M.....	16	180
St. Joseph, Mo.....	61	670
Springfield, Mo.....	46	420
Texarkana, Ark.....	35	190
Waco, Tex.....	50	460

The unit of observation here is the city, and each city has two *paired* numerical values—the size of its labor force and the annual sales in its Excelsior store. The statistical problem is to discover whether there is a correspondence of movement, that is, a “going-togetherness”—whether size of labor force and amount of annual sales move together. This varying-together of the series of paired items is called *concomitant variation*.

Statistical inquiry into concomitant variation is correlation analysis. Thus, in correlation we do not deal with one series but rather with the association or relationship between series, and we do not measure variation in one series but rather compare variation in two or more series. Do the series vary together in the same direction; do they vary together in opposite directions; or do they not vary together at all?

To measure the degree of association between the movements within two or more series, it is not necessary that there be proportionate change; that is, it is not necessary that for every

increase in the size of labor force in each city there be always a proportionate increase in department-store sales. It is sufficient for the series of paired items to manifest concomitance of movement on the average.

To measure the association of series through correlation we must have a sufficiency of items in the series. If we have only two or three pairs of values, we cannot *generalize* concerning the way in which they vary together. We must have either all * the information that exists or else enough of it to be representative. Moreover, there must not be a blank in one series where there is a value in the other series; there must be pairing throughout.

To measure the degree of relationship between two or more variables, we must have grounds *other than statistical* for holding that the variables are interconnected in nature or society. If they are interconnected, we want to know to what degree.

Correlation and Causation

The existence of correlation between variables does not necessarily mean that one is the cause of the movement in the other. For example, there is a high degree of correlation between the hardness of a metal and its tensile strength (the harder the metal the greater the tensile strength), but that correlation does not mean that hardness causes tensile strength. The cause may lie entirely elsewhere.

The relationship of hardness and tensile strength is functional; one depends upon the other. "Functional" means that the two attributes are interdependent. But in order to establish functional relationship, we have to go beyond the confines of statistical analysis to other fields of knowledge. We must do likewise to establish causal relationship.

All by itself, correlation does not manifest causation or functional relationship. By itself it establishes only *covariation*. Even

* Of course, if the complete information consists of only a few pairs of values, then the results of correlation analysis may not be valid. Moreover, if certain of our observations are extreme values, then the relationship may be distorted, as in information about New York City lumped with corresponding information about 10 small towns.

in the case of spurious correlation, there is covariation. But what is spurious correlation?

Spurious Correlation

There may be covariation where there is no logical basis for relationship. A measure of correlation may be arrived at on the basis of covariation, but it may be nonsensical or without meaning. In short, it is spurious.

Two aspects of spurious correlation are worth brief discussion.

1. The covariation may be coincidental only. That is, there may be concomitant variation between two variables when the two variables do not operate in the same physical or social systems, when they have nothing to do with each other. Professor Harold Larrabee * gives instances of coincidental correlation between variables, such as the increase in the number of inmates of institutions for the insane during the years 1920 to 1931 which ran parallel with the unprecedented increase in college enrollments during that period; or the close correspondence between the average salary over time of Presbyterian ministers in Massachusetts and the price of rum in Havana, Cuba.

The number of such instances can be multiplied. They all result from dragging together two things that cannot possibly go together in the world as we know it.

2. The covariation between two variables may really be due to the action of a third variable on the other two. Changes in sales in 5-and-10-cent stores concomitant with changes in stock-exchange prices may not be functionally or causally related; both may depend upon a third variable, namely cyclical fluctuations.

The Scatter Diagram

The first step in correlation analysis is to visualize the relationship.

For each unit of observation in correlation analysis there is a pair of numerical values. One is considered the independent

* *Reliable Knowledge*, p. 368, Houghton Mifflin Co., 1945.

variable; the other is considered dependent upon it and is called the dependent variable. Thus, sales may be analyzed as dependent upon advertising expenditures, post-office receipts upon population, industrial accidents upon length of working day. In some cases, the relation of dependence is not easily ascertainable. In such cases, we may call that variable dependent which follows later in time. If the time sequence, too, is not clear, an arbitrary assignment of dependence must be made.

Every unit of observation—in our Excelsior case, each city—is represented by a dot on a graph. This dot simultaneously shows the value of the independent variable on the X -axis and the value of the dependent variable on the Y -axis. When every dot representing a pair of numerical values has been plotted, we have what is called a *scatter diagram*, exemplified in Chart 20.1 for the data in Table 20.1. The term “scatter” refers to the dispersion of the dots on the graph.

The general rules for graphic presentation (see Chapter 5) hold for the scatter diagram with certain exceptions. There need not be a zero origin on the Y -axis; thus, both axes start so as to accommodate the lowest values in the series. In a scatter diagram the dots are not connected. Sometimes the dots are labeled.

In plotting a dot, we find the value of the X variable and plot it against its paired value; thus in Chart 20.1, for Abilene, Texas, we plot a dot at 22 on the X -axis against 250 on the Y -axis, using the paired values as shown in Table 20.1.

Types of Relationship

The dots in a scatter diagram may fall in a line; then there would be no scatter at all. Or the dots may be scattered over the graph without any system. Or they may scatter systematically; that is, they may form a path. The way the dots scatter gives us an indication of the kind of relationship which exists between the two variables.

If, generally, low values of one variable go with low values of the other variable and high values go with high values, then we have a path that runs roughly from the lower left corner to the

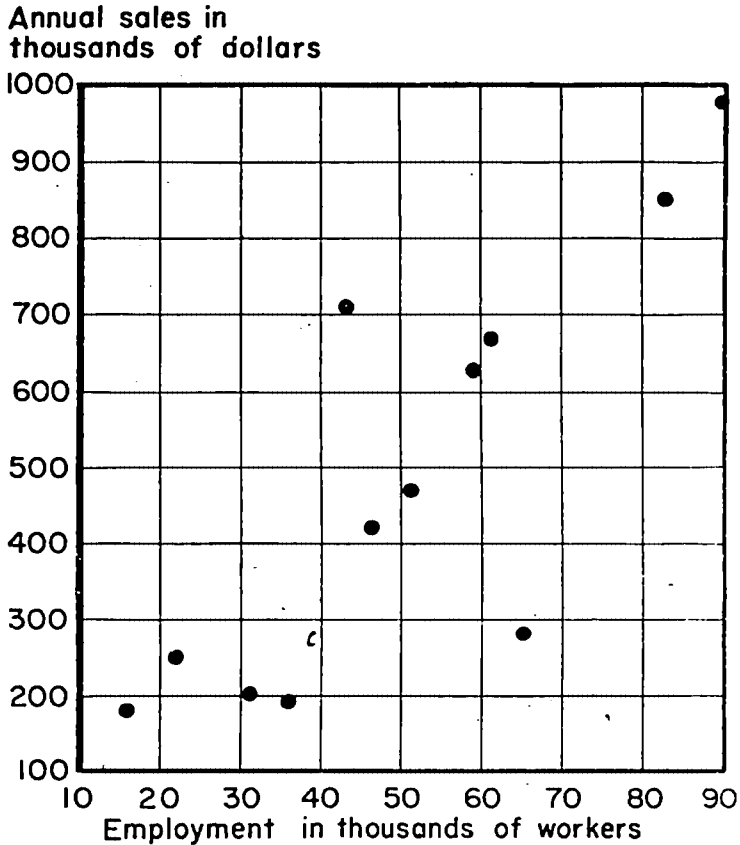


Chart 20.1. Scatter Diagram of 1950 Sales of the Excelsior Department Stores and Estimated Employment Figures in 12 Cities.

Source: Table 20.1.

upper right corner as shown in Chart 20.1. Here low sales generally go with small employed labor force, and high sales generally go with large employed labor force. This relationship is direct and is termed *positive* correlation.

On the other hand, if high values of one variable go with low values of the other variable, the path runs roughly from the upper left corner to the lower right corner, as shown in Chart 20.2. High productive efficiency is generally concomitant with

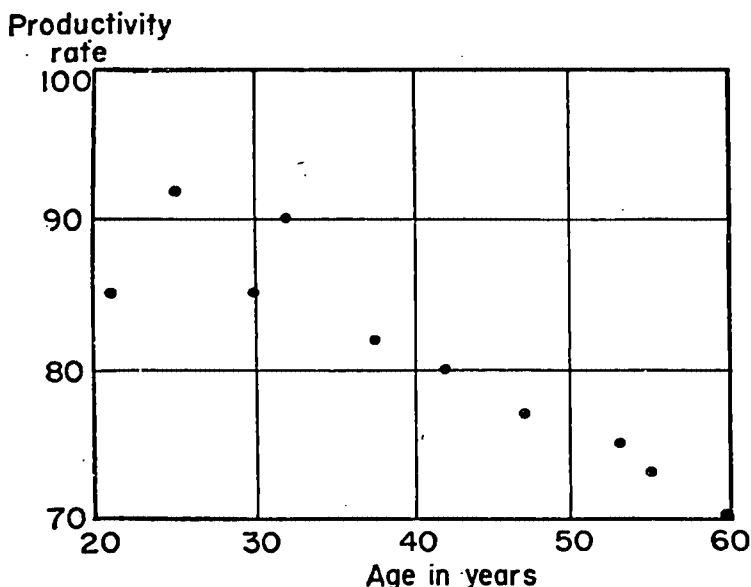


Chart 20.2. Age and Productivity Rate for 10 Selected Workers in the Transcontinental Smelting Corporation of Omaha, Nebraska, December 1, 1957.

low age, and low productive efficiency with high age, on a certain type of job. This relationship is inverse and is termed *negative* correlation.

As an ideal case, all the dots may form a line rather than merely a path. Such a situation, never found in practical work, constitutes *perfect* correlation. Perfect correlation might be either positive or negative.

If no path is discernible from the dots, then there is no association between the variables; the independent variable has no influence on the dependent variable, as shown in Chart 20.3. There is no relationship between the weight of a student and his grade in statistics. This lack of relationship—where one variable does not help us to establish the value of the other—is termed *no correlation*.

Thus, after having some practice and gaining some familiarity with correlation, we can gauge with some accuracy the

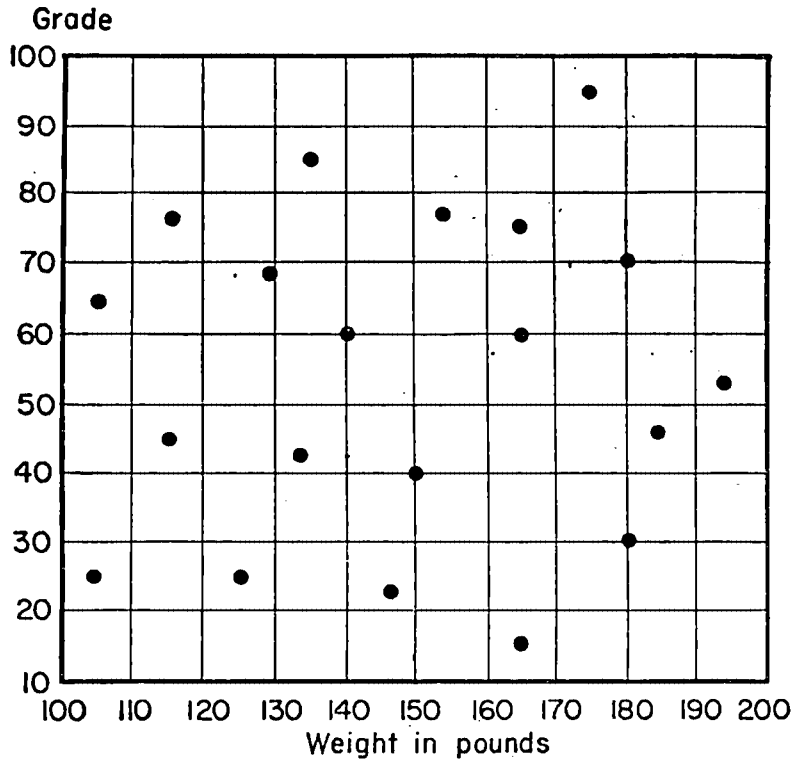


Chart 20.3. Weight and Final Grade in Statistics of 20 Students at a College in the United States, January 1956.

extent of relationship between two variables and the type of relationship from the scatter diagram. But the exact degree of relationship cannot be obtained from the scatter diagram.

BASIC CONCEPTS

In the light of the preceding discussion we are ready to develop a method of correlation analysis. We will devise measures of correlation, developing formulas for them in general terms. For this discussion we will sometimes need to use numerical measures referring to our example of the Excelsior Department Stores. But

the actual computation of these measures will not be undertaken until the end of the chapter, when relatively easy methods will be presented, methods simpler than the development formulas.

The Regression Line

If the dots on a scatter diagram follow a path, this path may be described by a line.* Since the dots by themselves do not form a line, this line describes only the *average* path of the dots. In our example, it would not describe the relationship of sales to employment *in any given city* but rather the *average* relationship between the two variables for the whole series. Thus, it is a line describing the relationship between the two variables on the average and may be called "the line of average relationship." This line may be thought of as a dynamic mean, and is analogous to the concept of the mean of a series.

This concept of the line of average relationship was first used by Sir Francis Galton in a study of hereditary characteristics, where he found that the height of descendants has the tendency to *regress* toward the average height of the race. This tendency led Galton to call this line of average relationship the *line of regression*. This latter term is still widely used, although "regression" has no meaning in terms of the data in problems ordinarily under investigation in the fields of economics and business.

The regression line is analogous to a trend line in a time series, where there is a functional relationship between time and other data, as sales or production or employment. Therefore, we can compute the regression line in the same way that we computed the trend line. The best method would be that of least squares with the setting up of two normal equations. This approach leads us to understanding and interpretation of the line of regression. For purposes of measurement we can use a short method to be discussed later in this chapter.

The major practical use of the regression line is as an *estimating line*. The estimating line is fitted by means of an *estimating equa-*

* This line may be straight or curved. For the present we deal with straight lines; we mention curved lines later (pages 373-374).

tion, $Y_c = a + bX$. The equation is also called the *regression equation*. For any given value of X , the independent variable, we can find an estimated value of Y , the dependent variable, by use of the estimating equation. For any X value, whether occurring or not occurring within the range of the series, we can estimate a corresponding Y value. Thus, if the Excelsior Department Stores want to open a new store in Mill City with its employed labor force of 100,000 and want to get an estimate of sales to be expected in Mill City, the Excelsior statistician can arrive at the estimate by means of the estimating equation, in which he substitutes 100,000 for X . Hence the X variable is called the "predicting" variable and the Y variable the "predicted" variable.*

The equation for the regression line (which will be computed later) is, as we know, $Y_c = a + bX$. This is the equation used in finding straight-line trend in time-series analysis (Chapter 15). If $a = -20$ and $b = 10.1$, and if X , the employed labor force in Mill City, is 100, then Y_c is \$990. Since the figures are in thousands, the estimated annual sales for the new store are \$990,000. (See page 373.)

Standard Error of Estimate

Just as we found it necessary to supplement an average for a series with a measure of dispersion or variation, so in order to see how good a representative the regression line is as a description of the average relationship between the two series, we look for a measure of dispersion about it. If we have a wide scatter or variation of the dots about this regression line, then it would have to be considered a poor representative of the relationship. The more closely the dots cluster around the line, the more representative it is and the better the estimates based on the equation for this line. And if the dots should all lie on the regression line (a hypothetical situation), then there is no variation about

* In this usage, "predicting" and "predicted" do not always refer to the future; we often predict a Y value occurring at the same time as its corresponding X value.

It must be emphasized that predicting a Y value from an X value beyond the range of the given observations should be done cautiously since the relationship may be operative only within the limits of the observations.

the line and the correlation is perfect. This last ideal situation, if it occurred with regard to sales and size of employed labor force, would mean that sales in a city could be *exactly* predicted from the size of the employed labor force in that city. This situation is not realistic, since factors other than employment also influence department-store sales. These other factors pull some or all of the dots away from the regression line.

This variation about the line of average relationship can be measured in a manner analogous to the measuring of the variation of the items about an average. Thus, we use here a measure of variation similar to the standard deviation—the *standard error of estimate*.* It is computed as is a standard deviation, being also a square root of the mean of squared deviations. But the deviations here are not the deviations of the items from the arithmetic mean; they are rather the vertical distances of every dot from the line of average relationship, as visualized in Illustration 20.1.

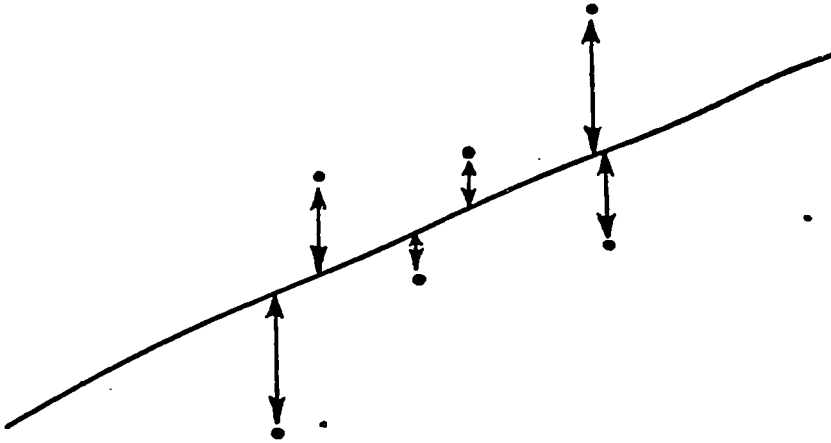


Illustration 20.1. Deviation of Dots from Their Line of Average Relationship, as Basis for Measurement of Their Dispersion through the Standard Error of Estimate. Each dot indicates a Y value; each corresponding point where the arrow meets the line indicates a Y_0 value.

* The standard error of estimate is the same *concept* as the standard deviation; this fact does not mean that it always has the same *value* as the standard deviation. They measure different things by the same method.

The deviation of each dot from the regression line is symbolized by $Y - Y_c$. Thus, the square root of the mean of the squared deviations is

$$S_Y = \sqrt{\frac{\sum(Y - Y_c)^2}{N}}$$

with S_Y standing for the standard error of estimate. We will later in this chapter discuss the computation of the standard error of estimate by a short method.

Moreover, the standard error of estimate is interpreted analogously to the standard deviation. Thus, within a band of $\pm 1S_Y$ from the regression line, 68.27% of the dots fall; within a band of $\pm 2S_Y$, 95.45% of the dots fall; and within a band of $\pm 3S_Y$, 99.73% of the dots fall.* Thus, within a band of $\pm 3S_Y$, we will find practically all of the dots. These bands are shown schematically in Illustration 20.2.

If we estimate a Y value for some X value, using the estimating equation or regression equation, the standard error of estimate puts us in a position to determine how accurate our estimate is. We already know that for Mill City, with an employed labor force of 100,000, the Excelsior Department Stores arrived at an estimate of annual sales of \$990,000. Since this is an estimate, the board of directors may well ask, "How good is this estimate?"

Their statistician bases his answer upon the standard error of estimate, as follows: The estimated sales are a point on the regression line. The sales that will really eventuate after the store is open are represented by a dot somewhere on the scatter diagram. Where is this dot with regard to its estimated counterpart on the regression line? The chances are about 2 out of 3 (or 68.27 out of 100) that the actual sales will turn out to be within the $1S_Y$ bands, that is, the estimated sales $\pm 1S_Y$, or \$990,000 $\pm$ \$148,430,† or between \$841,570 and \$1,138,430. The chances are

* These percentages hold true if the distances of the dots from the regression line form a "normal distribution." If the distances are not "normally distributed," then the percentages of the dots within the limits correspond *approximately* to these percentages.

† See computation on p. 370.

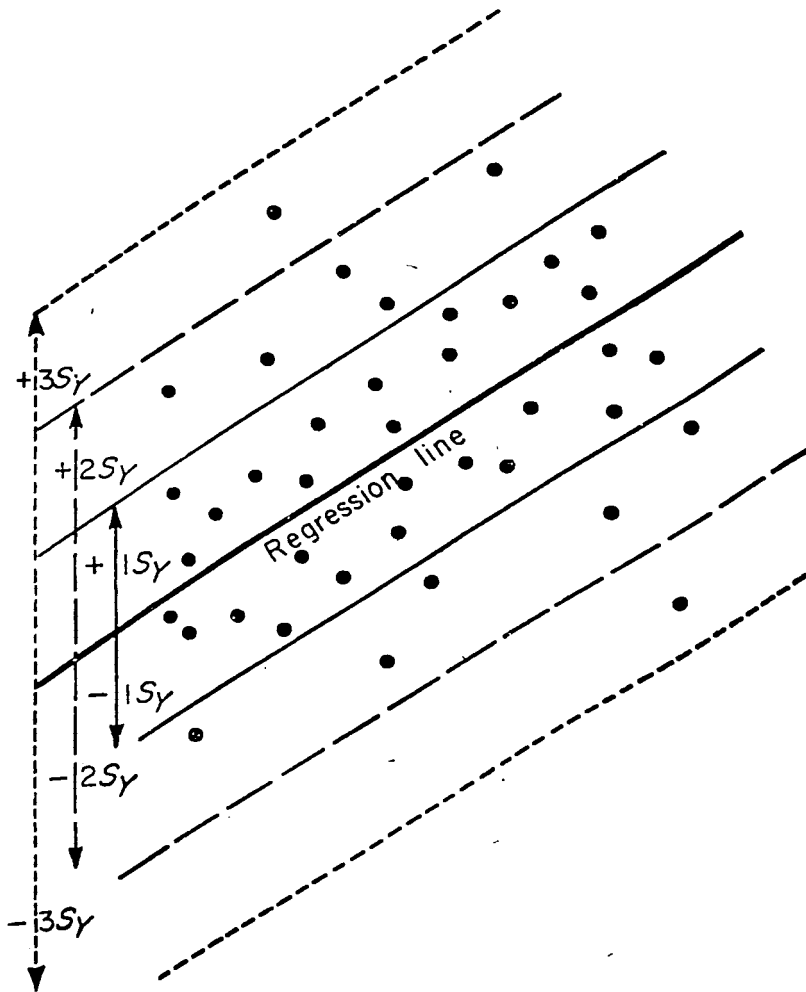


Illustration 20.2. Standard Error of Estimate Bands about Regression Line.

about 21 out of 22 (or 95.45 out of 100) that the actual sales will turn out to be $\$990,000 \pm \$296,860$ or between $\$693,140$ and $\$1,286,860$. It is practically certain (the chances are about 369 out of 370, that is, 99.73 out of 100) that the actual sales will turn out to be $\$990,000 \pm \$445,290$ or between $\$544,710$ and $\$1,435,290$.

Coefficient of Correlation

The standard error of estimate measures the scatter of the dots about the regression line. If the scatter is large, S_Y is large and the relationship is small. Thus, S_Y is a measure of the relationship between the variables, but it is an absolute measure expressed in the same units used to measure the dependent variable Y . If Y in two different correlation problems is in different units, however, we cannot compare the two relationships in terms of S_Y . We might, for example, have established: (1) the relationship between sales and employment, wherein Y and S_Y are numbers of dollars; and (2) the relationship between union membership and employment, wherein Y and S_Y are numbers of union members. These are units of different kinds and we cannot compare them. Moreover, the size of the possible variation in one series may be far larger than the size of the possible variation in the other series. Therefore, to compare measures of relationship we must be able to express the *degree of association* in an abstract form. This can be done by finding a *coefficient of correlation*.

This coefficient can be arrived at by expressing the standard error of estimate S_Y relative to the standard deviation s_Y .* This comparison of S_Y and s_Y is the key to measuring the degree of relationship divorced from units of Y .

The standard error of estimate of the Y values can be very small compared with the standard deviation of the Y values. In fact, it can be zero. A zero value for S_Y would mean that all dots are on the line. Sales in that case would thus be entirely accounted for by the size of the employed labor force. This condition would be perfect correlation.

The standard error of estimate can be large compared with the standard deviation; in fact, it can be as large as the standard deviation. If $S_Y = s_Y$, there is "no correlation"; there is a wide scatter of the dots—they follow no path since any X value may

* s_Y is the standard deviation of the Y values and is computed like any standard deviation. (See Chapter 13.)

go with low, medium, and high Y values. An example is shown in Chart 20.3; weight of students does not guide the dots onto a path and weight is therefore of no help at all in explaining grades in statistics. Since any value of X may be paired with high, low, and medium values of Y , the regression line in this case of "no correlation" runs parallel to the X -axis. This condition means that variations in X fail to explain variations in Y , and explain only themselves.* Hence, Y is independent of X . In Illustration 20.3 it is shown that $S_Y = s_Y$ in the case of "no correlation."

In between these two extremes of $S_Y = 0$ (perfect correlation) and $S_Y = s_Y$ (no correlation), there are degrees of relationship based upon a comparison of S_Y with s_Y . How large S_Y is, compared with s_Y , shows the degree of correlation.

This comparison leads us to a measure of the degree of relationship between the two variables. This measure could be symbolized by S_Y/s_Y , or preferably S_Y^2/s_Y^2 . But in this form it would lead to illogical results, since in the case of perfect correlation ($S_Y = 0$), the measure would be 0, and in the case of "no correlation" ($S_Y = s_Y$) the measure would be 1.† But perfect correlation means unity of movement, and therefore a measure of such unity should give us the figure 1. Complete absence of relationship should give us 0. To achieve a measure which will conform to this logic, we subtract S_Y^2/s_Y^2 from 1. We take the square root of this subtraction and obtain Karl Pearson's coefficient of correlation, symbolized by r . The formula is

$$r = \sqrt{1 - \frac{S_Y^2}{s_Y^2}}$$

We attach to the result a plus or minus sign obtained from inspection of the slope of the regression line. We shall later see that r can be more easily found.

In Pearson's formula, "no correlation" means $r = 0$ and per-

* In such a case, to predict a Y value we would use $\bar{Y}$ as an estimate since X does not help us.

† Actually, S_Y^2/s_Y^2 measures the inability to explain Y through X (see p. 365). Thus, 1 would mean "perfect failure" to explain the variation in Y , or "no correlation."

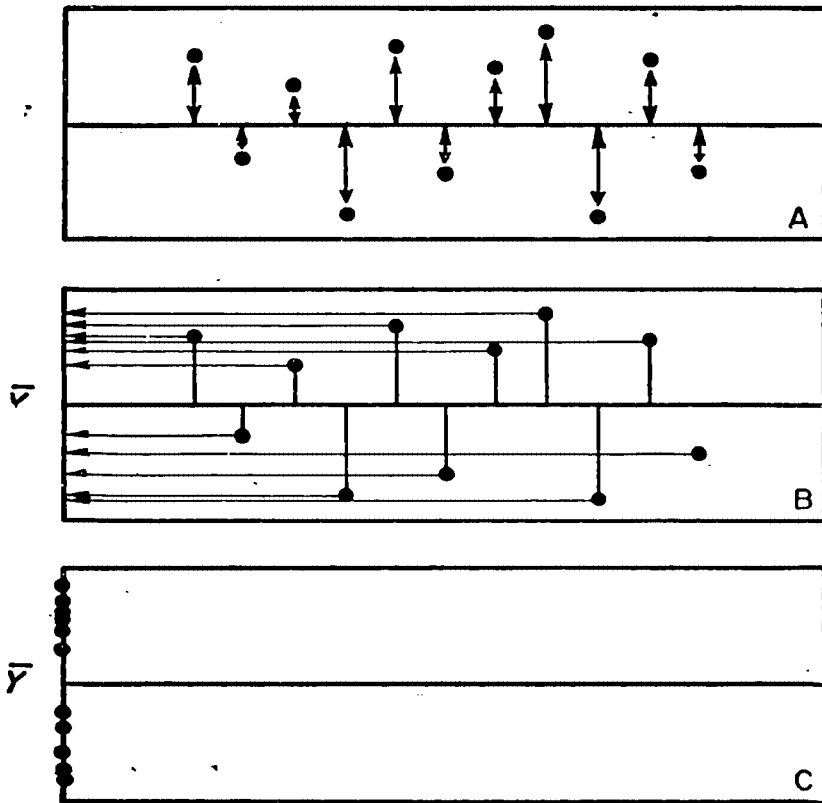


Illustration 20.3. $\bar{S}_Y = s_Y$ in Case of "No Correlation."

A. S_Y Based on Vertical Distances of Dots from Regression Line.

B. Dots Moved to $\bar{Y}$ -Axis.

C. Dispersion of Dots around $\bar{Y}$ as Basis for s_Y .

The dispersion in C around $\bar{Y}$ is identical with the dispersion around the regression line in A.

fect correlation means $r = 1$. Perfect correlation can be positive or negative. Thus, perfect positive correlation is represented by $+1$ and perfect negative correlation is represented by -1 . These are the limits of the measure called the coefficient of correlation. The closer two variables come to unity of movement, direct or inverse, the closer they come to being represented by the figure 1, either positive or negative. Consequently, in measuring correla-

tion, it is not possible to get a figure higher than 1. In practice, such unity is never found. Measures of correlation take such form as $+.95$, $+.37$, $+.15$, or $-.90$, $-.32$, $-.06$. Thus, a measure of correlation can be anywhere from $+1$ to -1 , with 0 meaning no correlation.

The coefficient of correlation r is an abstract number or a pure number; that is, it is not expressed in units of a variable like number of dollars or number of union members. It is an abstract number which measures the degree of the relationship between two variables.

When we compute the coefficient of correlation for sales of Excelsior Department Stores in relation to the size of the employed labor force in the cities where they are located, we shall find (on page 369) that $r = +.83$. This betokens high positive correlation.

Coefficient of Determination

Of the variation in sales in the Excelsior Department Stores, there is a part which is explained by size of the labor force in the cities, and a part which is not. There are three different kinds of variation in the problem. (1) The variation of the Y values, or actual sales figures, from their mean; (2) the variation of the dots from the regression line; (3) the variation of the computed Y values from their mean.

1. The first kind of variation is represented by s_Y , the standard deviation of the Y values, in our case the sales data. This variation measures the differences among sales in different cities of the Excelsior chain. These differences are a resultant of *all* the forces which bring about variation in these sales data. Thus the measure, s_Y , may be looked upon as the measure of the *total* variation. But part of this total variation among sales data has been explained and part has been left unexplained. What has been explained and what has not been explained?

2. The *unexplained* variation is the variation of the dots from the regression line. The dots are pulled away from the regression line by factors other than size of labor force. The size of the

labor force is not the only determinant of sales. This dispersion of the dots brought about by factors other than size of labor force is measured by S_Y , as we already know. Thus, S_Y may be looked upon as the measure of unexplained variation.

3. We are already familiar with s_Y and S_Y . But there is another measure of variation which is of importance in this context. This is the measure of *explained* variation. This measure in our example shows how variation in size of labor force affects size of sales. The explained variation is based upon the deviation of the Y_c values from the arithmetic mean of all the Y_c values. The Y_c values in our example are the sales values computed by means of the regression equation as opposed to the Y values which are the original sales values. The measure of the variation of the Y_c values from their mean* is σ_{Y_c} .

An important relationship can be established which holds true only for the squares of the measures s_Y , S_Y , and σ_{Y_c} . s_Y is, as we have just said, the total variation; s_Y^2 is the total variance, since, as was mentioned in Chapter 13, the square of the standard deviation is the variance.

The total variance is the sum of the unexplained and the explained variance, or in symbols,

$$s_Y^2 = S_Y^2 + \sigma_{Y_c}^2.$$

In case of perfect correlation, S_Y is zero, as we have already seen. Therefore, from this formula it is clear that in this case the total variance is explained variance. In case of "no correlation," $S_Y = s_Y$; therefore the total variance is unexplained variance. In this latter case, the variation of the computed Y values around their mean, σ_{Y_c} , is zero because all the dots are on a horizontal regression line and every dot has the same value as $\bar{Y}$.

S_Y^2 is the unexplained variance. To measure how much of the total variance is unexplained, we compare the unexplained variance S_Y^2 with the total variance s_Y^2 , that is, we compute

* The mean of the computed values, or $\bar{Y}_c$, equals the mean of the Y values or $\bar{Y}$. This is so because $\Sigma Y_c = \Sigma Y$ (see the third mathematical characteristic of the mean, p. 141).

S_Y^2 divided by s_Y^2 . This gives a coefficient which is called the *coefficient of nondetermination*. This coefficient is given the symbol k^2 and is expressed in percent. In our problem, the coefficient of nondetermination is 31%. This means that of the total variance in sales of the Excelsior Department Stores we fail to explain 31% through the size of the labor force, as can be seen below.

σ_{Yc}^2 is the explained variance. To measure how much of the total variance is explained, we compare the explained variance σ_{Yc}^2 with the total variance s_Y^2 , that is, we compute σ_{Yc}^2 divided by s_Y^2 . This also gives a coefficient; but this coefficient is called the *coefficient of determination*.

We can put this coefficient of determination in more convenient form, and in terms already familiar. The formula showing the relation of the total variance and its two parts is, as we have just seen,

$$s_Y^2 = S_Y^2 + \sigma_{Yc}^2.$$

If we divide every expression in this equation by s_Y^2 , we obtain

$$1 = \frac{S_Y^2}{s_Y^2} + \frac{\sigma_{Yc}^2}{s_Y^2}.$$

Since S_Y^2/s_Y^2 is the coefficient of nondetermination and σ_{Yc}^2/s_Y^2 is the coefficient of determination, therefore we may now say that the coefficient of nondetermination plus the coefficient of determination is equal to 1 (or 100%).

Hence, the coefficient of determination may be expressed as 1 minus the coefficient of nondetermination. In symbols, this is

$$\frac{\sigma_{Yc}^2}{s_Y^2} = 1 - \frac{S_Y^2}{s_Y^2}.$$

The right-hand expression is familiar to us from Pearson's coefficient of correlation, which, we saw, is

$$r = \sqrt{1 - \frac{S_Y^2}{s_Y^2}}.$$

Hence, the coefficient of determination can now be expressed as the square of r .*

Thus, to find the coefficient of determination, we merely have to square r . In our case of sales and size of labor force in cities, r^2 is 69%. This means that 69% of the variation in sales has been explained by the variation in size of labor force.

The coefficient of determination tells us how successful we have been in explaining sales through size of labor force in these 12 cities. Thus the coefficient of determination helps us to decide whether it is necessary to look for additional determinants of sales, as, for example, income level, commercial and industrial character of the communities, number of competing stores in these cities.

COMPUTATION OF MEASURES

We have discussed the meaning and formulas for computing the basic concepts of correlation. These formulas are needed for understanding and interpreting the various concepts since they bring out clearly their essential nature. We can, however, arrive at the same results by simpler computational methods. These simpler methods are not suitable for comprehension of the basic concepts. But now that the basic concepts have been presented, we may employ these simpler methods.

There is more than one approach to finding measures of correlation. We shall use what is called the product-moment approach.† In our approach, it is necessary to start with the computation of r . In finding r we also obtain all information required to compute the standard error of estimate and the equation for the line of regression.

* Since the coefficient of nondetermination is k^2 , then $r^2 + k^2 = 1$. Hence, the coefficient of nondetermination can be obtained by subtracting r^2 (in percent) from unity (100%).

† See the Appendix to this chapter for the method of rank correlation.

Computation of r

In measuring variation in one series, we based a measure of variation upon the distance of each item from the mean ($X - \bar{X} = x$ for some one series and $Y - \bar{Y} = y$ for another series). In correlation analysis, we are interested not in variation, but in covariation. Here each observation is not one item in a single series, but a pair of items from two series. Therefore, the raw material is not x or y but the two combined, or xy . This product signifies "co-deviation," and we may employ as a measure of "co-deviation" the mean co-deviation or $\Sigma xy/N$.

If there is little or no association between the two series, then X may go with small or large values of Y . That is, for any x the corresponding y may be positive or negative. In that case, the products xy will tend to cancel out, and their sum Σxy will be negligible. If Σxy comes out positive, the association is direct or positive; if Σxy comes out negative, the association is inverse or negative.

This formula $\Sigma xy/N$ may not be adequate, if the series are in different units. Therefore, to express correlation in terms divorced from units we express the x and y in terms of their respective standard deviations. This procedure leads to the formula

$$r = \frac{\Sigma \left(\frac{x}{s_x} \cdot \frac{y}{s_y} \right)}{N}$$

This is a basic formula for the coefficient of correlation.

Several manipulations are possible which make this formula easier to use. In our computations, we employ the following:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$$

In Table 20.2, the coefficient of correlation for sales in Excelsior Department Stores and size of employed labor force in cities with such stores is worked out for the data in Table 20.1. The coefficient of correlation r is found to be $+.83$.

TABLE 20.2. COMPUTATION OF THE COEFFICIENT OF CORRELATION FOR DATA IN TABLE 20.1.

Employment, thousands, X	Sales, thousands of dollars, Y	$x =$ $X - \bar{X}$	$y =$ $Y - \bar{Y}$	xy	x^2	y^2
22	250	-28	-235	+ 6,580	784	55,225
31	200	-19	-285	+ 5,415	361	81,225
90	980	+40	+495	+19,800	1,600	245,025
82	850	+32	+365	+11,680	1,024	133,225
43	710	- 7	+225	- 1,575	49	50,625
65	280	+15	-205	- 3,075	225	42,025
59	630	+ 9	+145	+ 1,305	81	21,025
16	180	-34	-305	+10,370	1,156	93,025
61	670	+11	+185	+ 2,035	121	34,225
46	420	- 4	- 65	+ 260	16	4,225
35	190	-15	-295	+ 4,425	225	87,025
50	460	0	- 25	0	0	625
600	5,820	0	0	+57,220	5,642	847,500
ΣX	ΣY	Σx	Σy	Σxy	Σx^2	Σy^2

$$\bar{X} = 50, \quad \bar{Y} = 485$$

$$\begin{aligned}
 r &= \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}} = \frac{+57,220}{\sqrt{5642 \times 847,500}} = \frac{+57,220}{\sqrt{4,781,595,000}} \\
 &= \frac{+57,220}{69,149} = +0.83
 \end{aligned}$$

Computation of S_Y

The standard error of estimate can be found as a by-product of the computation of r .

Pearson's formula for r (page 362) is

$$r = \sqrt{1 - \frac{S_Y^2}{s_Y^2}}$$

We now know r , and s_Y can easily be found from the computation for r in Table 20.2. Since $s_Y = \sqrt{\Sigma y^2 / N}$, it is in our case $\sqrt{847,500 / 12}$. Therefore, $s_Y = 266$. We may now use Pearson's

formula to find S_Y . For convenience, we bring S_Y to the left side of the equation, resulting in the following:

$$S_Y = s_Y \sqrt{1 - r^2}.$$

Thus, in our case of sales and labor force,

$$S_Y = 266 \sqrt{1 - (.83)^2} = 148.43,$$

which being in units of thousands of dollars is the figure \$148,430 that was mentioned in discussing the standard error of estimate (page 359).

Computation of the Regression Line

The computation of the equation for the regression line likewise can be simplified. Our goal is to fit a straight line to depict the path of the dots. Such a line can be fitted through the method of least squares. We wish, therefore, finally to obtain an equation of the type $Y_c = a + bX$, where a tells us how high above the X -axis the regression line is started and b tells us how much the line increases or decreases for each unit of X .

Thus we may use the two normal equations to find a and b , and then we can find from the basic equation a Y_c value for every X . The two normal equations, as we know (see pages 359–360) are:

$$\begin{aligned}\Sigma Y &= Na + b \Sigma X, \\ \Sigma XY &= a \Sigma X + b \Sigma X^2.\end{aligned}$$

Fitting a straight line to the scatter is simplified considerably if we take the point of averages ($\bar{X}$, $\bar{Y}$) as the point of origin. It will be recalled that in the least-squares short method for trend (Chapter 15, pages 248–253) the first normal equation, $a = \Sigma Y/N$, states that $a = \bar{Y}$ at $\bar{X}$ (with $\bar{X} = 0$). In trend fitting, we moved only the Y -axis (over to the point of $\bar{X}$). Here in correlation we likewise take $\bar{Y}$ to be 0, and therefore also move the X -axis (up to $\bar{Y}$) as shown in Illustration 20.4.

Since the regression line now runs through the point of origin, the line does not start above the X -axis but at the X -axis; and therefore there is no Y intercept and $a = 0$.

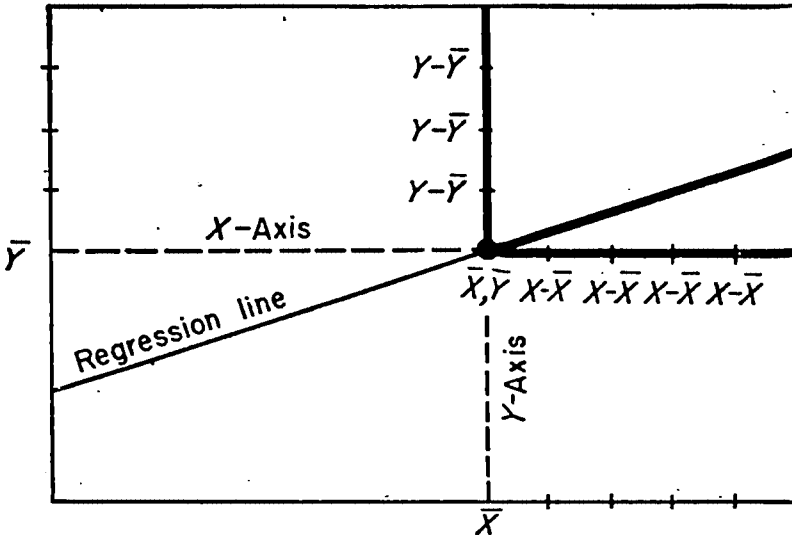


Illustration 20.4. Basis for Computing the Regression Line.

The X and Y origins are at the mean of X and the mean of Y respectively. Thus distances from the origin become distances from the mean. But distance of an item from the mean, $X - \bar{X}$, is x . Analogously, $Y - \bar{Y} = y$.

Hence, the equation $Y_c = a + bX$ now becomes

$$y_c = bx.$$

Since this equation is in terms of deviations rather than original values, it is inconvenient to use, though correct. With very little effort, we can transform it so as to make it more convenient to use.

Since $y = Y - \bar{Y}$ and $x = X - \bar{X}$, the equation $y_c = bx$ becomes

$$Y_c - \bar{Y} = b(X - \bar{X}).$$

We know $\bar{Y}$ (which is the same as $\bar{Y}_c$) and we know $\bar{X}$; we can obtain b as a by-product of having found r . Based on the second normal equation in the short method that we discussed in trend analysis (Chapter 15, page 249), $b = \Sigma XY / \Sigma X^2$. Here, however, we are employing deviations and the formula is in terms of x, y ,

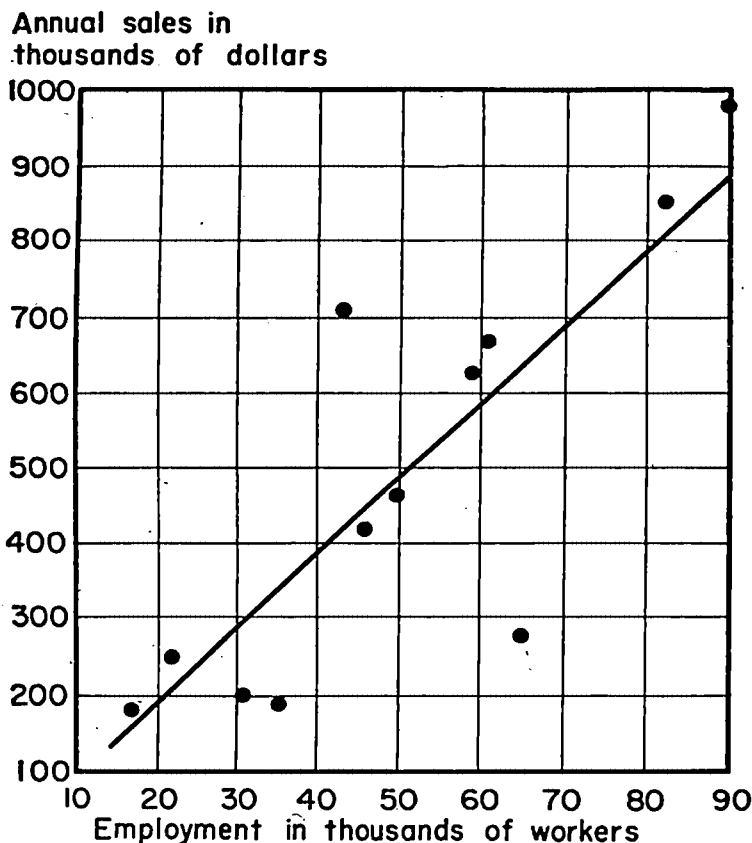


Chart 20.4. Regression Line and Scatter Diagram for Data in Table 20.1.

and x^2 rather than X , Y , and X^2 , or $\Sigma xy/\Sigma x^2$. The work sheet for finding r (Table 20.2) gives us Σxy and Σx^2 , as well as $\bar{X}$ and $\bar{Y}$. In our problem of the relation of sales and employment, $b = 57,220/5642 = 10.1$, $\bar{X}$ is 50, and $\bar{Y}$ is 485. Therefore, the equation for the regression line here is

$$Y_c - 485 = 10.1(X - 50),$$

or

$$\begin{aligned} Y_c &= 485 + 10.1(X - 50) \\ &= -20 + 10.1(X). \end{aligned}$$

This equation has an origin of $X = 0$. In practical terms, such an origin does not always make sense. In our case it would mean a city with no employment, and the origin must obviously here be considered a hypothetical point. In our equations above, a is -20 at the X origin.

Now we can substitute original X values in our equation instead of x values, and use the equation for plotting the regression line by finding Y_c values. Of course, this straight line can be plotted by finding any two Y_c values. Chart 20.4 shows the regression line fitted on the scatter diagram.

The regression equation is especially valuable to estimate or predict Y values corresponding to given X values, based upon their average relationship. In practice, we can set up as a valuable summarization of our results a predicting table:

<i>Employment in Thousands of Workers</i>	<i>Predicted Annual Sales in Thousands of Dollars</i>
50	485
100	990
150	1495

SCOPE OF CORRELATION

In our previous discussion throughout this chapter, two conditions were present: (1) There were only two variables, one independent and one dependent; (2) the relationship could best be described by a straight line. The first condition results in *simple correlation*; the second results in *linear correlation*. Thus we have been discussing *simple linear correlation*. Moreover, we have restricted ourselves to ungrouped data.

As distinct from linear correlation, there is a type known as curvilinear correlation. Here the path of dots shows one or more bends. There may be a high positive correlation between sales performance and years of experience of salesmen, for instance, but a point is reached where sales performance declines after

age becomes a limiting factor. In economics, supply and demand schedules in many cases are illustrative of curvilinear correlation.

As distinct from simple correlation, there are two types, one known as multiple correlation, the other as partial correlation. Multiple correlation involves measurement of the degree of association between more than two variables, with one variable dependent and the other variables independent. It would necessitate multiple correlation, for example, if we sought to measure the combined influence on sales of size of labor force, educational level, and commercial and industrial characteristics of the city.

In multiple correlation, we would consider the combined effect of more than one determinant on sales. If we seek to discover just the influence of employment on sales while taking into consideration the fact that other variables are also operative, then we are measuring partial correlation.

Summary

1. Correlation analysis deals with the association between two or more variables.
2. Statistical measures of correlation are proof only of co-variation between series, not of functional or causal relationship.
3. The scatter diagram is a special type of graph which gives some indication of the type and degree of relationship between variables.
4. The relationship between series may be direct (positive), inverse (negative), or lacking.
5. The basic concepts in correlation analysis are the regression line, the standard error of estimate, the coefficient of correlation, the coefficient of determination.
6. In the product-moment approach used here, first r is computed, then the other measures.

7. Rank correlation is discussed in an Appendix to this chapter.

8. Correlation may be simple or multiple or partial. It may be linear or curvilinear.

APPENDIX—RANK CORRELATION

In Chapter 20 of this book the coefficient of correlation was found by the product-moment method. Another method of measuring the degree of association between two series is known as rank correlation. Its principal uses are three: (1) Rank correlation may be used as an alternative in finding the degree of association for series of data originally given in measurements, such as size of labor force in certain cities and sales data for the Excelsior Department Stores in these cities. Rank correlation for a small number of paired items is a quick, nonlaborious estimate of association which may be useful where intensive study is not called for or where we wish to discover speedily whether we should go to the trouble of finding r . (2) Rank correlation is used to find the degree of association for series of data that are given as ranks, scores, or standings, such as the standings of the same group of students in two different examinations. (3) Rank correlation permits us to correlate two sets of *qualitative* observations which are subject to ranking, such as the qualitative productivity ratings (poor, fair, good, superior) given a group of workers by two independent observers.

Let us illustrate how a coefficient of rank correlation is found where the two series of data are given as measurements, as in the Excelsior Stores example in Chapter 20. Each X —here the size of labor force in the city—is ranked in the X series in order of its value; and each Y —here the sales volume of the Excelsior store in the same city—is ranked in the Y series * in order of its value. The ranking in both

* If in either series one or more values of X or Y occur more than once then we give each of these repeated values the same rank based on the mean of the ranks they jointly occupy. Thus in the series 32, 79, 82, 91, 91, 91, 130, the value of 91 occupies the fourth, fifth and sixth ranks. Since the mean of 4, 5, and 6 is 5, we give each value of 91 a rank of 5. Sometimes joint ranks will be in fractions as in the series 23, 32, 48, 48, 56, 89, where the two values of 48 fall in the third and fourth ranks. The mean of 3 and 4 is 3.5; therefore the rank value of 48 in each case is 3.5.

series may be in ascending or in descending order of value in the original data. Here let us rank the data on page 349 in ascending order as follows:

<i>Size of Labor Force</i> (X)	<i>Sales Volume</i> (Y)	<i>Rank of X</i>	<i>Rank of Y</i>
16	180	1	1
22	250	2	4
31	200	3	3
35	190	4	2
43	710	5	10
46	420	6	6
50	460	7	7
59	630	8	8
61	670	9	9
65	280	10	5
82	850	11	11
90	980	12	12

If each X and its paired Y have exactly the same rank we have perfect positive correlation with a coefficient of $+1$. If, on the other hand, the ranks are reversed in the two series so that the highest-ranking X goes with the lowest-ranking Y and so on, we have perfect negative correlation with a coefficient of -1 . A glance at the ranking of the Excelsior Department Stores data just above shows that the correlation will not be perfect since neither of these two extreme possibilities is present.

A measure of rank correlation which has been in wide use was developed by Charles Edward Spearman. Spearman's coefficient of rank correlation bears the symbol ρ ,* the Greek letter *rho*, as distinguished from the product-moment coefficient which is symbolized by the letter r .† The formula for Spearman's coefficient of rank correlation is

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)},$$

* Some statisticians have used the symbol r_r for Spearman's coefficient of rank correlation; some the symbol r_{rank} ; some the symbol ρ_r .

† More recently a method of rank correlation has been developed by Maurice G. Kendall in *Rank Correlation Methods* (London, Charles Griffin, 1948). Kendall's coefficient bears the symbol τ (the Greek letter *tau*).

where ρ stands for the coefficient of rank correlation,
 D stands for the difference between each pair of ranks,
 Σ is the familiar symbol for "sum of," and
 N stands for the number of paired items.

To find D for each pair of ranks we subtract the rank of Y from the rank of its corresponding X .

Below are the paired ranks of the Excelsior Stores data with their corresponding differences D and the squares of these differences.

<i>Rank of X</i>	<i>Rank of Y</i>	<i>D</i>	<i>D</i> ²
1	1	0	0
2	4	-2	4
3	3	0	0
4	2	2	4
5	10	-5	25
6	6	0	0
7	7	0	0
8	8	0	0
9	9	0	0
10	5	5	25
11	11	0	0
12	12	0	0
			58

Substituting in the formula for rank correlation, we obtain

$$\begin{aligned}\rho &= 1 - \frac{6(58)}{12(144 - 1)} = 1 - \frac{348}{1716} \\ &= +0.80\end{aligned}$$

The coefficient of rank correlation of +0.80 in this example is not identical with, but comes close to, the coefficient of correlation of +0.83 found for the original data by the product-moment method on page 369.

If there are extreme values in the series the method of rank correlation serves to de-emphasize these extremes by giving them a *rank* in order of value as opposed to the product-moment method where computation is based on *actual* values. But Spearman's coefficient of rank correlation should not be indiscriminately taken as a substitute for the coefficient of correlation computed by the product-moment method, since it often is only a rough measure of association.

PART
4

Inductive Statistics

Elements of Sampling Theory

Sampling in Everyday Living

Every day, people in the ordinary business of life make judgments and take action based on samples. From casual conversation with the mail man, the butcher, the automobile mechanic, and a neighbor, a person draws conclusions concerning the opinion of the "public." This type of thinking from a few particulars to a generality is characteristic of sampling procedure. Daily life is full of such instances—the housewife who tastes a grape before she buys a pound, the teacher who asks a student a few questions to determine his grasp of an entire subject. People are not usually aware that such a pattern of thinking is sampling.

A grocer receives from a new supplier a shipment of oranges, consisting of several crates. He opens *some* of the crates, selects *several* oranges from each crate, and examines them. Since the oranges examined are acceptable, he accepts the entire shipment. Later, after he has sold a number of oranges, he asks *some* of the customers who bought them how they liked them. Most of the customers asked are satisfied and make favorable comments; the grocer assumes that the new supplier's oranges meet with general approval.

Before he places another order with this new supplier, however, the grocer compares the sample of oranges with a sample taken from crates received from the old supplier. He finds the sample of the new supplier's oranges considerably better. Finally he compares the new sample with what he considers standards for quality, for instance weight, of oranges of this type; he finds that the new oranges come close to these standards. He then decides to order more oranges from this new supplier.

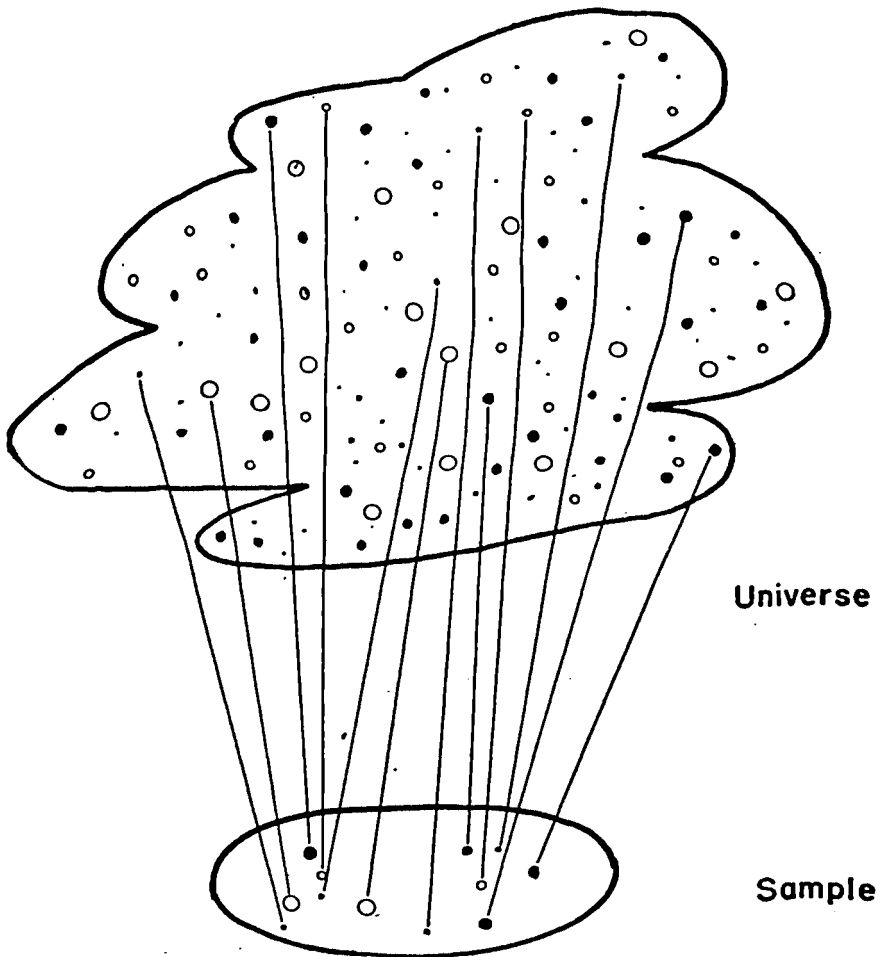


Illustration 21.1. Schematic Diagram of a Sample.

Our grocer took a number of samples: one when he examined *some* of the oranges delivered to him by the new supplier, one when he interviewed *some* of the customers who bought this fruit, and one when he examined *some* of the oranges from the old supplier. The sample in each instance was part of a large group. In each instance he applied the findings based on the sample to the larger group from which it was taken. Thus, the samples are used as "examples," and in fact the word *sample* derives from the Latin *exemplum*, which means *example*.

The grocer's way of taking samples may be called the rule-of-thumb method. If we introduce the scientific approach, the rules of thumb give way to statistical rules, and we substitute statistical method for the hit-or-miss method.

The concept of a sample is shown schematically in Illustration 21.1. The larger group from which the sample is drawn is called the *universe* or the *population*.*

Induction

Reasoning from sample to universe is called statistical induction. In logic, induction refers to an important method of gaining knowledge. Inductive reasoning involves observing individual cases and arguing from them to a general principle. In statistical induction we make inferences concerning the characteristics of a large group, sometimes of an infinitely large group, from the characteristics of a specifically chosen, usually small, part of the group that we observe. We draw conclusions from the observed smaller group and apply these conclusions to a larger mass.

All inductive reasoning involves generalizing. In statistical induction we are able to add to our generalizations a statement as to their precision. After we have defined the trait to be measured, statistical induction then involves: (1) measuring this trait in a sample; (2) applying the measures to the universe from which the

* The meaning of the term "population" in statistical sampling differs from the ordinary meaning of this term. In statistics "population" is not restricted to mean a number of persons. For example, in statistics we also speak of a population of prices.

sample was taken; (3) estimating the error involved in this application.

THE UNIVERSE AND THE SAMPLE

A universe or population may be defined as an aggregate of items possessing a common trait or traits. This definition points up two factors: a universe is the complete group of items about which knowledge is sought; each and every item has some certain specified attribute or attributes. Thus in our illustration about the grocer, *all oranges received from the new supplier* was a universe.

A universe such as all the oranges from this new supplier contains a known number of items and is called *finite*. On the other hand, an *infinite* universe is a universe containing a limitless number of items; for instance, the universe of all possible opinions from every point of view of all consumers about a certain commodity.* In many instances a universe is finite but large. We do not always know the precise size of a finite universe although theoretically its size is ascertainable. In such instances the universe is, in practice, treated as if it were infinite.

The grocer's sampling process may also serve as an illustration of another distinction between universes. In his study of the customers' opinions he took samples of a *human population*. In choosing the oranges to be examined, on the other hand, he was taking samples of an *inanimate population*. Studies of human populations are typical of problems encountered in market research and demography, whereas studies of inanimate populations are typical of industrial quality control† and accounting. In addition, samples may be taken from an *animal population*, as in biological studies.

Each universe consists of items or elements which we call *sampling units*. For example, in our instance of the grocer an

* Further discussion of finite and infinite universes will be found on p. 411.

† An Appendix to Chapter 23 deals more extensively with statistical quality control.

orange is the sampling unit. This concept must be distinguished clearly from that of the *characteristic being investigated*. While oranges are the sampling units, their weight may be the characteristic being investigated.

A sample for the purposes of statistical induction, as outlined on page 382 of this chapter, must consist of two or more sampling units. If oranges are the sampling units, one orange is not sufficient for these purposes.*

A sample is that part of the universe which we select for the purpose of investigation. A sample should exhibit the characteristics of the universe; it should be a "microcosm," a word which literally means "small universe."

There are various types of samples. In this chapter we shall deal only with one major class of samples, the *random sample*.† Randomness is based upon the concept of probability and sampling theory is applicable only to such a sample. In addition to randomness, size is another aspect of a sample that is of importance. Statisticians distinguish between a "large" and a "small" sample. By "small" sample is commonly understood any sample that includes 30 items or fewer; for example, a sample of 12 oranges is a "small" sample.

Why a Sample Is Used

The study of a sample offers several advantages over a complete investigation of the universe. Why and when is it desirable to study a sample in preference to the universe?

In some cases an appraisal of a sample is the only way of making a study. In testing materials the quality of the item often can be measured only by destroying it. For example, to ascertain the maximum pressure that glass containers can withstand, we put a sample of these containers under pressure until they break. For such destructive testing the use of samples

* A sample consisting of only one sampling unit may yield information but we cannot estimate the error involved in the application of the sample measure to the universe.

† Detailed discussion of random samples will be found on pages 431-445.

is obviously the only feasible method, since we certainly do not wish to destroy every item we produce.

But even when investigating a sample is not the only way, it is often the best. Complete enumeration, which means investigating every item in the universe and is the alternative to investigating a sample, is sometimes practically impossible and is always much more expensive and time consuming. It can readily be seen, for instance, that in establishing a brand preference for soap it would cost too much money and consume too much time to ask every person in the United States. Even if complete enumeration can be done once, it may prove too costly in money and time if investigation has to be repeated, as it must be in periodic market research to keep check on consumer preference. Moreover, complete enumeration may take so much time that the findings are of no use by the time they become available.

Furthermore, study of a sample may be qualitatively superior in its results. There will always be some error in every survey result whether it is obtained from a sample or by complete enumeration. This error is the difference between the obtained and the true values. The *total* error in a given study consists of several component parts; the error introduced by the use of a sample is only one of these parts. In many studies based on every item in the universe the error component introduced by interviewers, checkers, and other such agents is very large. For work with a sample, fewer people are needed, and consequently better qualified people can be obtained and they can be given better training. Since better and better-trained interviewers can be used, this error component may be much smaller in a sample study. Thus the error introduced by not having every item investigated may be outweighed by the reduction of this other error component, and the *total* error may be lower in the sample study than in a study based on every item in the universe.

But apart from these considerations, complete enumeration is not necessary since sound sampling procedures will yield satisfactory results. Of course we cannot expect from a sample

exactly the same answer that might be had from complete enumeration, but the statistician can establish limits on the basis of sample results within which—with a known probability of being right—the true answer lies.

Thus, sampling usually offers great advantages as to cost and time, and sometimes even as to quality of results. In practical work, sampling frequently means being able to make a study which otherwise would not be feasible. "When we are able to generalize logically and precisely about thousands of values which we have never seen, simply upon the evidence afforded by, say, fifty or a hundred of those values, we have come a long way indeed toward an intellectual mastery of the world around us." *

SAMPLE-UNIVERSE RELATIONSHIPS

Concepts of Estimation and Statistical Significance

Thus far we have emphasized basic concepts involved in statistical induction. In statistical work we come upon three main problems of the relation of the sample and the universe. They are:

1. When we have measured a trait in a sample, what can we say about this trait in the universe?
2. Does our sample come from a given universe?
3. Does our sample come from the same universe as some other given sample?

Let us clarify these three problems, in order, through three sets of illustrative cases.

1. Based on evidence obtained for some year from a *sample* of persons living in the United States, we can estimate the average annual income of *all* persons in the United States. Or from a *sample* of orders received by a large mail-order house during some week we can estimate the percentage of *all* orders

* Roy W. Jastram, *Elements of Statistical Inference*, California Book Co., Ltd., Berkeley, California, 1947, p. a of Preface.

that were received from a certain territory during this week. These are problems in statistical *estimation*.

2. We may know from official sources the average wage of *all* production workers in the rainwear industry in the United States for 1955. If we establish the average wage for a *sample* of production workers in the rainwear industry in 1956, we can ascertain whether the 1956 sample might "come from" the 1955 universe of wages. Certain further conclusions may be drawn from the answer to this question. This is a problem in statistical *significance of a difference*.

In the example of rainwear wages we investigated the relation between a sample measure and a *known* universe measure. The same problem can be investigated for a *hypothetical* or *assumed* universe value. For example, we may assume that 10% of women living in the United States use a certain cosmetic. This figure of 10% may be based on past experience and knowledge of the market. A sample of women living in the United States may show a different percentage of users of the cosmetic. We can establish whether this sample might or might not have come from the universe for which we assumed the figure of 10%. This is another type of problem in statistical *significance of a difference*.

3. We have two samples of monthly rentals, each sample drawn from rentals in a different part of the same city. The mean rentals in these two samples differ. We can ascertain whether the two samples could come from one universe or whether they could come from different universes. If they come from different universes, we may be able to call their difference in mean rental "real," that is, not due to chance. This is a problem in statistical *significance of a difference between the means of two samples*.

Another illustration: in a large insurance company we select a sample of key-punch operators on the day shift and a sample on the night shift. For each of these two samples we establish the percentage of operators who make punching errors. The percentage of the operators making punching errors is different for each sample. Can the two samples come from the same

universe? If it is unlikely that they come from the same universe, then the difference in the percentage of operators on the two shifts making punching errors may be "real," that is, not due to chance. This is a problem in statistical *significance of a difference between percentages in two samples*.

Detailed discussion of statistical estimation will be found on pages 413-420 and of statistical significance on pages 420-429.

THE THREE DISTRIBUTIONS

The practical illustrations of problems in estimation and statistical significance presented in the preceding discussion are indicative of the vast range of such problems that face statisticians. Sampling theory has made it possible to deal effectively with these problems.

We turn then to a consideration of the rudiments of sampling theory. Let us consider data consisting of all expense vouchers of the Magnum Company, a very large corporation, for one year. We are interested in problems such as the mean or median value of the vouchers, their variability, the percentage of vouchers relating to the expenses of executives.

Let us illustrate sampling theory first through the problem of the mean voucher value.

From the standpoint of sampling theory we are concerned with three different frequency distributions, their characteristics and relations. These three frequency distributions may be called: (1) the universe distribution; (2) the sample distribution; (3) the sampling distribution. These we now discuss, define, and explain.

The Universe Distribution

Let us assume that we have investigated the universe and have full knowledge of its mean and its standard deviation. Although this assumption is rarely practicable in actual investigations, it will be helpful in acquiring an understanding of sampling theory.

A study of the universe, in this case all expense vouchers, reveals the distribution of voucher values shown in Illustration 21.2. We symbolize the *mean of the universe*, which is also called

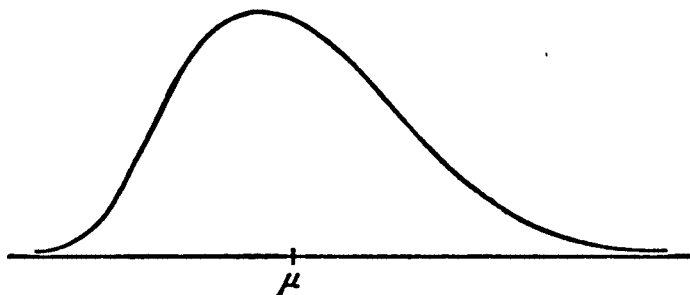


Illustration 21.2. Distribution of Expense-Voucher Values for the Magnum Company.

the *true mean*, by μ (the lower-case *m* in Greek and called *mu*) and its standard deviation by σ (the lower-case *sigma*). Greek letters are used for these measures to emphasize their difference from corresponding measures taken from a sample. Measures characterizing a universe, such as μ and σ , are called *parameters*.

The shape of the distribution of expense-voucher values shows slight skewness but does not depart radically from a normal distribution. It is not necessary for the universe to be normally distributed for sampling theory to apply.

The Sample Distribution

From this universe of expense vouchers of the Magnum Company for some one year, let us take a random sample* of 100 vouchers and analyze it. Its distribution is shown in Illustration 21.3. A sample distribution may take any shape. We designate the summary measures of the sample distribution by symbols already familiar; thus $\bar{X}$ is used for the sample mean and s for

* A random sample is a sample so selected from a universe that each item has an equal or known chance of being included in it. Detailed discussion of random sampling will be found on pages 431-445.

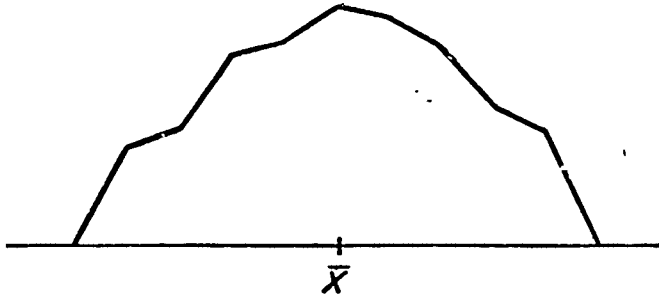


Illustration 21.3. Sample Distribution of 100 Expense Vouchers, the Magnum Company.

the standard deviation of the sample. A measure characterizing a sample, such as $\bar{X}$ or s , is called a *statistic*.

Let us imagine (in practical work it will not be done) that we have drawn and measured a large number of random samples of 100 vouchers each. (See Illustration 21.4.) The mean values of such voucher samples can be expected to differ somewhat from one another. These sample means constitute the raw material out of which a sampling distribution is constructed.

The Sampling Distribution

Let us take the means of the many samples suggested in Illustration 21.4 and construct a frequency distribution from them. Such a distribution of sample means is called a *sampling distribution* (see Illustration 21.5).^{*} Each item in such a distribution is the mean of a sample. It is important to note that sampling distributions of means closely approximate a normal distribution,[†] and it has been established that the mean of a sampling distribution is the same as the mean of the universe from which the samples were taken. Thus the mean of a sampling distribu-

^{*} Here we restrict our discussion to the sampling distribution of *means*. We may have sampling distributions of other measures, such as the median, the standard deviation, and the like.

[†] If the universe distribution is not normal, then the sampling distribution of sample means approaches normality as the sample size increases.

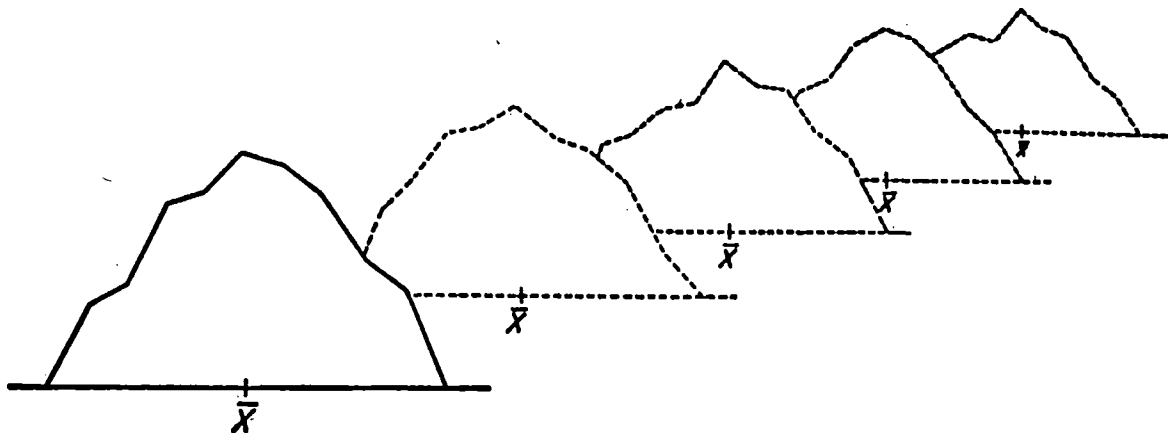


Illustration 21.4. Some of the Large Number of Sample Distributions of 100 Expense Vouchers, the Magnum Company

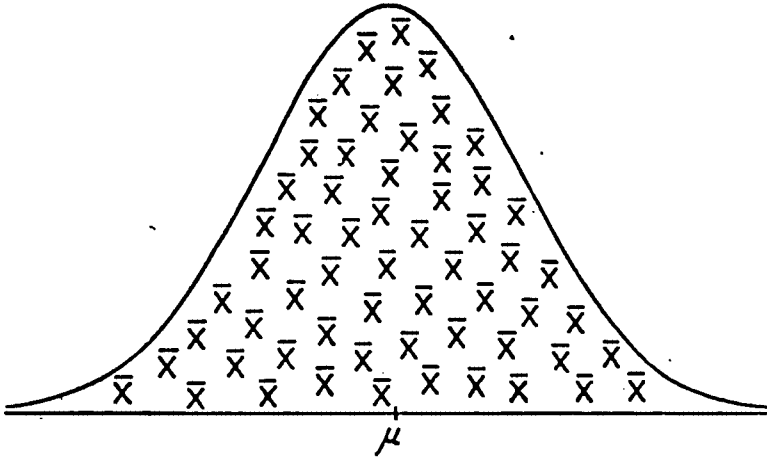


Illustration 21.5. Sampling Distribution of Sample Means Based on a Large Number of Sample Means.

tion may be designated by the same symbol as that for the mean of the universe involved, namely μ . The standard deviation of the sampling distribution is given a special name, *standard error of the mean*, and is symbolized by $\sigma_{\bar{x}}$. The subscript indicates that in this case we are dealing with a sampling distribution of means.

Relations of the Three Distributions

Starting with the description of a sample, let us make use of the concept of sampling distribution in order to arrive at statements about characteristics of the universe. Illustration 21.6 depicts the flow of this thinking.

A consideration of the means and standard deviations of the three distributions helps make clear the relations of these distributions.

MEANS. We know that it has been established that the mean of the sampling distribution is the same as the mean of the universe.

Many sample means will differ from the universe mean. We

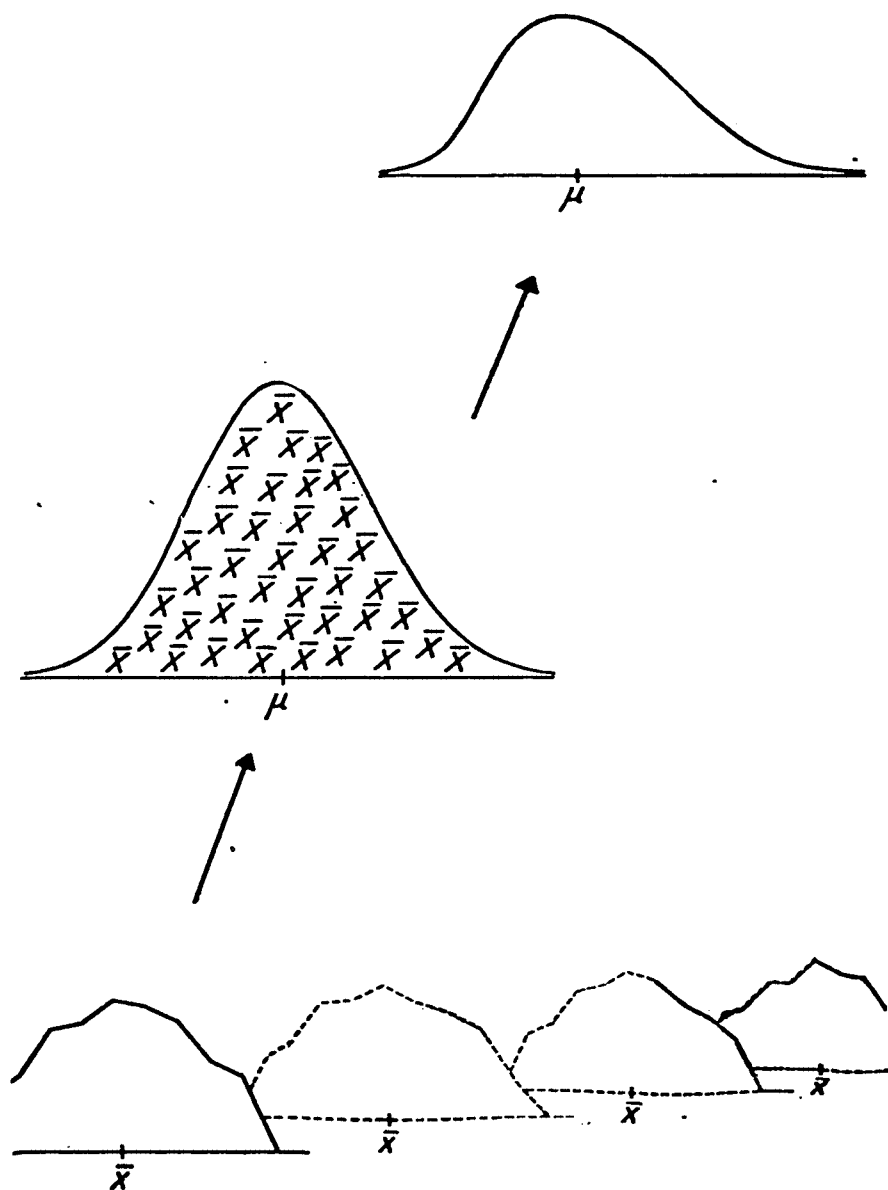


Illustration 21.6. The Three Distributions—a Schematic Presentation: Sample Distributions $\rightarrow$ Sampling Distribution $\rightarrow$ Universe Distribution.

cannot expect to hit exactly upon the mean value of a large universe of expense vouchers every time we find the mean value of a sample of 100 vouchers. But we may be able to use the sample information as an estimate of the universe value. We will return to this last point on page 414.

STANDARD DEVIATIONS. The observed standard deviation of a sample has been found to be close to the standard deviation of the universe values. In fact, the standard deviation of the sample is usually so good an approximation that it may safely be used as an estimate of the corresponding universe measure. The formula for the standard deviation of a sample is $\sqrt{\Sigma x^2/n}$. This formula is the familiar one for s (see page 187) except that the lower-case letter n is used to indicate that the number of items refers to the sample size.

In order to use s of the sample to estimate σ of the universe we make a slight adjustment which has been found to contribute to greater accuracy of the estimate. This adjustment consists of using $n - 1$ instead of n . This adjustment decreases the denominator and therefore gives a larger result. Thus the estimated standard deviation of the universe is slightly larger than the observed standard deviation of the sample. This difference can be understood if we realize that the variability of all voucher values in the universe may be influenced by a few very large and very small values whereas it is not likely that this situation will occur in a sample.

We now know that the observed standard deviation of a sample may be used as an estimate of the universe standard deviation, with or without a slight adjustment.*

The third measure of variation is the standard deviation of the sampling distribution, known as the *standard error*. The standard error may be considered the key to sampling theory. The standard error of the mean can be found with the help of the universe standard deviation. It has been found that the standard error of the mean equals the standard deviation of the

* This qualification is made because the adjustment is often ignored in practical work, unless the sample is small.

universe divided by the square root of the sample size. The formula for the standard error of the mean is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}},$$

where $\sigma_{\bar{x}}$ is the standard error of the mean,

σ is the standard deviation of the universe, and

n is the number of items in the sample.

It is obvious that the standard error is smaller than the standard deviation of the universe since to get the standard error we divided the standard deviation of the universe by $\sqrt{n}$. This fact can also be made clear in another way. There may be some very large and very small voucher values in the universe. We may find a few extreme voucher values as small as \$1 and a few as large as \$500. This wide dispersion affects the magnitude of the standard deviation. But the standard error is a measure of variability in a distribution of *means*. If we draw equal-size samples, their means will differ but no mean will be \$1 and none will be \$500.

Thus a distribution of sample means is more uniform than the universe distribution. In fact, this greater uniformity is reflected in the formula for the standard error, which is based on the relation of the standard deviation of the universe and the sample size. The larger the sample size the smaller the standard error, and the smaller the sample size the larger the standard error. The standard error equals $\frac{1}{10}$ of the universe standard deviation if the sample size is 100 and $\frac{1}{20}$ if the sample size is 400 since the square root of 100 is 10 and the square root of 400 is 20.

An illustrative hypothesis will demonstrate this variation of the standard error. Suppose we took a large number of samples each the same size as the universe. Then the mean of each sample would be identical with the true mean. In such an extreme case, the "sampling distribution" would consist of means each of which is identical with the true mean and the standard error would be zero. Let us imagine for comparison a large number of

samples of, say, 50 items each. Some of these samples will have their means influenced by the chance inclusion of an undue number of small or large items which are not offset. Thus we get some sample means which are not close to the true mean. Such widening of the dispersion of sample means about the true mean is reflected in the larger standard error.

Moreover, it follows from the formula for the standard error that the larger the variation there is in the universe, the larger the variation in the distribution of sample means. Since there will be a larger dispersion in the universe of expense vouchers than, for example, in the universe of petty-cash vouchers, we accordingly expect a larger variability of sample means if we draw samples of expense vouchers than if we draw samples of petty-cash vouchers.

We based the formula for the standard error of the mean and discussions thereof on the standard deviation of the universe. Since the standard deviation of the universe is rarely known and since the very purpose of sampling is to avoid having to analyze the entire universe, we take for practical work the standard deviation of a sample which, as we already know, closely approximates the standard deviation of the universe. Consequently, to obtain the standard error for practical work it is sufficient to know only the measure of variation and the size of one single sample.

Thus, the formula for the standard error $\sigma_{\bar{x}}$ takes the following form:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

If perchance in practice we know several sample means from the sampling distribution, we may take the standard deviation of these several means from their mean as an approximation of the standard error.

We may now summarize the relations of the three distributions as far as means and standard deviations are concerned in the following table.

Measure	Distribution					
	Sample		Sampling		Universe	
	Symbol	Source	Symbol	Source	Symbol	Source
Mean	$\bar{X}$	Observed	μ	Estimated from $\bar{X}$	μ	Estimated from $\bar{X}$
Standard deviation	s	Observed	$\sigma_{\bar{X}}$	Estimated from $\frac{s}{\sqrt{n}}$	σ	Estimated from s

INTERPRETATION AND USES OF THE STANDARD ERROR OF THE MEAN

The standard error of the mean is defined as the standard deviation in the sampling distribution of means. It is a measure of the variation of sample means about the universe mean or true mean, and can be interpreted in the same way as a standard deviation of a normal distribution.

In Chapter 13 of this book the standard deviation was presented as a measure of the variation of items about the mean of the series, as in the cases of monthly rentals and of sales checks. Here in the sampling distribution the individual items are sample means and the measure of variation reflects sample-to-sample variation. This variation of sample means from the true mean—the “sampling error”—is due to the effect of chance in sampling. For example, if from a year's expense vouchers we draw five random samples of 100 each, we expect the means of these five expense-voucher samples to differ because of chance. These chance variations or sampling variations can be measured through the standard error.

Just as the standard deviation is used to locate an item within

a distribution, so the standard error of the mean tells us the location of a sample mean relative to the true mean.

Other things being equal, it is desirable to have a standard error as small as possible. The smaller the standard error the greater the uniformity of the sampling distribution and the better the chance that our single sample mean is located close by the true mean.

There are several very important practical uses of the standard error. Chapter 22 deals with this measure as it is used in *estimating* and in testing *statistical significance*. But at this point the standard error of the mean will be dealt with somewhat abstractly in order to deepen our understanding of basic concepts. To achieve this understanding let us discuss three situations in which the standard error is utilized.*

1. Sample mean known, true mean known or assumed.
2. Sample mean known, true mean unknown.
3. Some sample mean unknown, true mean known.

1. $\bar{X}$ KNOWN, μ KNOWN OR ASSUMED. The standard error can be used to specify the location of a known sample mean within the sampling distribution if the true mean is known or assumed. Illustration 21.7 shows the sample mean $\bar{X}$ as one standard

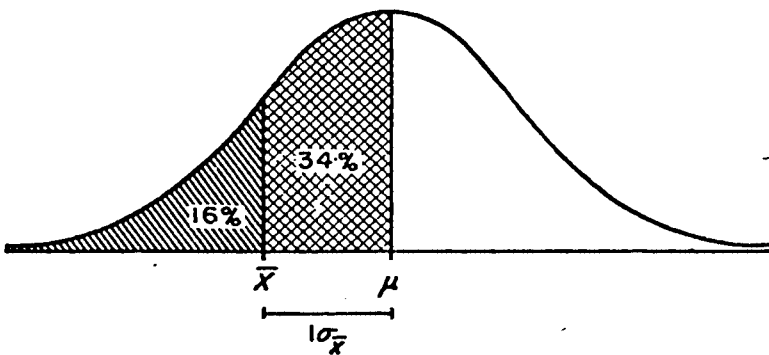


Illustration 21.7. Sample Mean One Standard Error Smaller than the True Mean.

* These three situations presume familiarity with areas under the normal curve; see pages 196-197 and pages 214-220.

error *smaller* than the true mean. This $\bar{X}$ being thus located, we know that about 34% of all sample means can be expected to have values ranging between μ and $\bar{X}$ while approximately 16% have a smaller value than $\bar{X}$.

Illustration 21.8 depicts the situation wherein a sample mean *exceeds* the true mean by $2\sigma_{\bar{X}}$.

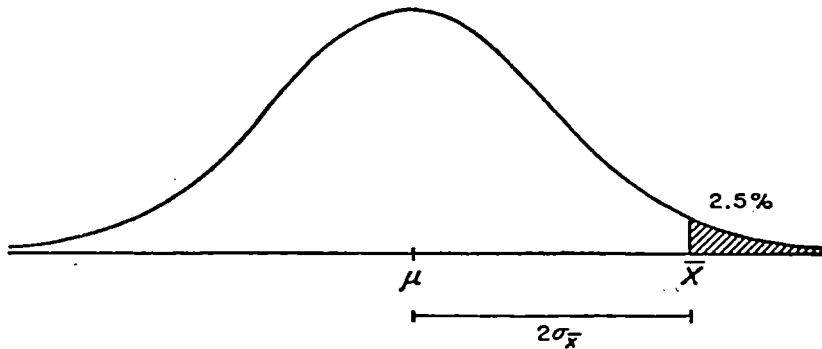


Illustration 21.8. Sample Mean Two Standard Errors Larger than the True Mean.

If a sample mean is located at a considerable distance from the true mean, we may consider the situation rare and call the sample which yielded such a distant mean a *rare sample*. The statistician has to designate which sample means he considers rare or unusual in a given situation. He may do so in terms of zones in the sampling distribution of the mean. In most situations the extreme 5% of sample means are called rare. But there are situations in which only "the extreme 1%" are considered rare.

If 5% is the criterion of the rare or unusual, we set up two areas in the sampling distribution which together encompass the extreme 5% of sample means by separating off $2\frac{1}{2}\%$ of the extremely low and $2\frac{1}{2}\%$ of the extremely high sample means. Illustration 21.9 depicts this situation. Accordingly we call every sample mean which falls into the marked-off areas in the tails of the curve a rare or unusual sample mean. Under this 5% definition of "rare," each rare sample mean is at least 2 standard errors away from the true mean.

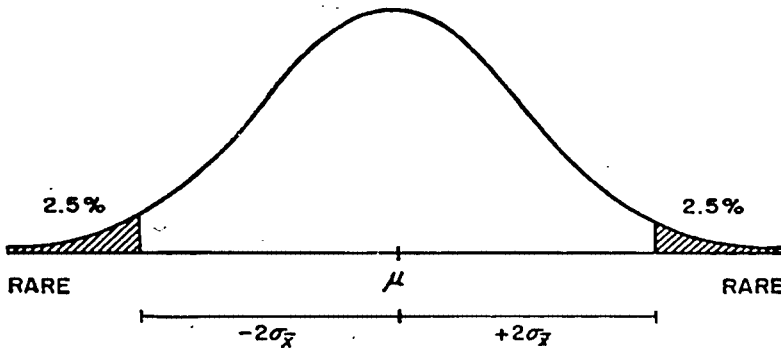


Illustration 21.9. Rare Sample Means (under the 5% Definition of "Rare").

2. $\bar{X}$ KNOWN, μ UNKNOWN. Any sample mean $\bar{X}$ must lie somewhere within its sampling distribution. But it may lie close to the center of the distribution or at an extreme end of it. The chances are very slight that any particular sample mean will fall *further* away than $3\sigma_{\bar{X}}$ from the unknown true mean. Therefore, adding $3\sigma_{\bar{X}}$ to a sample mean on each side makes us practically certain that the unknown true mean lies within this range. Here see Illustration 21.10.

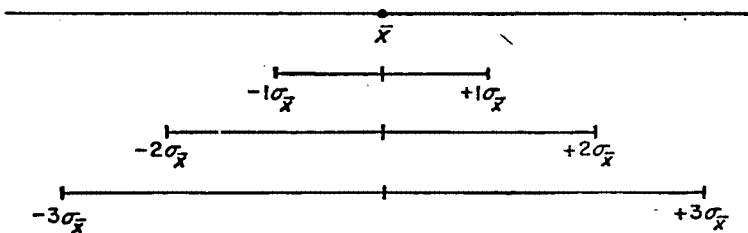


Illustration 21.10. Standard-Error Ranges about the Sample Mean (Sample Mean Known, True Mean Unknown).

3. SOME $\bar{X}$ UNKNOWN, μ KNOWN. The standard error of the mean, $\sigma_{\bar{X}}$, may be attached to the true mean or universe mean μ . Since the sampling distribution is normal, we know that of the sample means about 68% are found between $\pm 1\sigma_{\bar{X}}$ from μ , about 95% between $\pm 2\sigma_{\bar{X}}$, and more than 99% between $\pm 3\sigma_{\bar{X}}$.

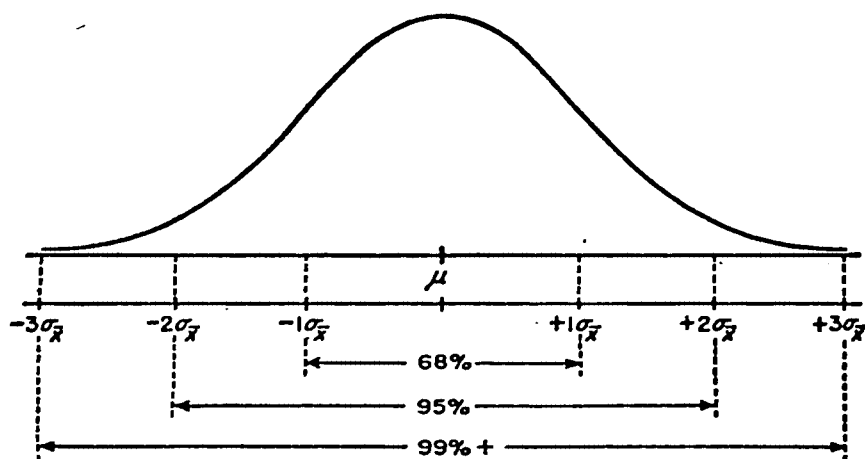


Illustration 21.11. True Mean and Standard-Error Ranges (Sample Mean Unknown, True Mean Known). Per cents are approximate.

(See Illustration 21.11). Such knowledge concerning these areas under the normal curve enables us to make statements about the probable location of any additional and yet unknown item—in this case, an *additional sample mean* derived from a random sample of the same size as the samples whose means form this sampling distribution. We may, for example, say that about 95 out of 100 *additional* sample means will fall within the range of $\pm 2\sigma_{\bar{x}}$ about the true mean.

STANDARD ERROR OF OTHER STATISTICS

Standard Error of the Median and of the Standard Deviation

In our study of expense vouchers, we may also be interested in the centrally located value and the variability, among other things. Hence we seek for the corresponding measures: the median (Med) and the standard deviation (s).

The concept of standard error is essential in these situations also.

Let us again envisage a very large number of random samples of equal size drawn from the universe of expense vouchers. The separate frequency distributions for all sample medians and for all sample standard deviations are the sampling distributions for the two statistics involved. If we have drawn a large number of samples, the sampling distribution of the median will approximate a normal distribution. This statement, however, only holds true where the *universe* does not depart greatly from normality. The sampling distribution of the standard deviation based on a large number of samples will approximate normality if the sample size is at least 100 and the universe is normal. The *true median* and the *true standard deviation*—that is, the two parameters—will be the mean of the sample medians and the mean of the sample standard deviations.

The measure of variation of the sample medians from their mean will constitute the standard error of the median. Likewise, the measure of variation of the sample standard deviations from their mean will constitute the standard error of the standard deviation.

The standard errors of these two measures—the median and the standard deviation—may be obtained readily from the standard error of the mean which has itself been obtained from the sampling distribution of sample means. It has been found that it is necessary merely to multiply the standard error of the mean by an adjustment factor in each case. Thus:

$$\begin{aligned}\sigma_{\text{Med.}} &= 1.25\sigma_{\bar{x}}, \\ \sigma_s &= .71\sigma_{\bar{x}},\end{aligned}$$

where $\sigma_{\text{Med.}}$ is the standard error of the median, and

σ_s is the standard error of the standard deviation.

The fact that the standard error of the median is approximately 25% greater than the standard error of the mean betokens that the median is subject to greater chance fluctuation. Thus if the median is not specifically indicated, the mean is preferable because of its smaller standard error. Consequently,

as far as sampling is concerned, we can express a preference for the mean in answer to the question asked on page 170 as to the appropriateness of the various averages.

The interpretation and use of the standard error of the median, and of the standard deviation (when the sample size is at least 100), are the same as that for the standard error of the mean.

Standard Error of the Total

Sometimes we are concerned with totals, for example the total value of all expense vouchers. The standard error of a total, σ_t , is equal to $N\sigma_x$ where N is the number of items in the universe, for example the number of all expense vouchers.

Standard Error of a Proportion

A situation which is frequently met in statistical work relates to data which fall into two categories. For example, workers may be male or female, punch cards may be acceptable or unacceptable, expense vouchers of a corporation may be made out by executives or nonexecutives.

Let us remain at the last illustration. An expense voucher would fall into one of two categories based on the possession or absence of an attribute, namely, being made out by an executive. What we are concerned to measure here is the proportion of vouchers in each category. Since the two proportions together constitute 100% of the vouchers, measurement of one will give us the other; thus, if 20% of all vouchers are from executives, 80% must be from nonexecutives.

To show how sampling theory applies to measurement of proportions, we here again envisage a sampling distribution. This distribution is based on a very large number of sample proportions, each symbolized by p . Each p stands for the proportion of "executive" vouchers in each of a large number of random samples of equal size. This sampling distribution of proportions approximates a normal distribution sufficiently to permit application of the concept. The concept of normal distribution here again is highly advantageous and is justified

unless, in the universe, the proportion under study is very large or very small *and* the sample size is small. The mean of this sampling distribution of proportions (see Illustration 21.12) is

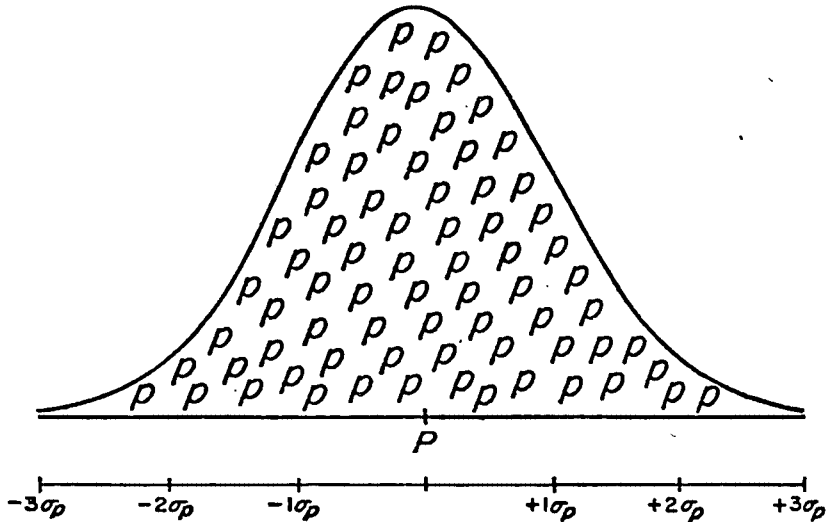


Illustration 21.12. Sampling Distribution Based upon Sample Proportions.

the true proportion, P , which is the same as the proportion in the universe. The variation in the sampling distribution is measured by the standard error of the proportion, σ_p .

The theoretical basis for the standard error of a proportion involves the true proportions of the two categories.

$$\sigma_p = \sqrt{\frac{PQ}{n}}$$

where σ_p = standard error of the proportion,

P = universe proportion, category 1,

Q = universe proportion, category 2, and $Q = 1 - P$.*

n = number of items in the sample.

P is the proportion of all expense vouchers made out by exec-

* 1 is used instead of 100% if proportions are expressed in their decimal form, as .25 instead of 25%.

utives and Q the proportion made out by nonexecutives of the corporation.

If the true proportion is unknown, it may be estimated from the proportion found in one sample. Thus the standard-error formula becomes

$$\sigma_p = \sqrt{\frac{pq}{n}}$$

where σ_p = standard error of the proportion,

p = sample proportion, category 1,

$q = 1 - p$ = sample proportion, category 2, and

n = number of items in the sample.

Standard Error of a Difference

Another problem of practical importance is that of differences between measures, for example, differences between means, differences between proportions. We may, perhaps, desire to compare the difference in the *mean* expense-voucher value in the Capital Steel Corporation and the Efficiency Steel Corporation. Or we may be interested in comparing the *proportion* of "executive" vouchers in the first-named corporation with the proportion in the second-named corporation.

Here we come upon a new measure, namely, the *difference* between means or proportions. Illustration 21.13 presents graphically the situation of the difference between means. Here we have two sampling distributions of means, sampling distribution #1 and #2. One distribution of sample means is based on samples drawn from Capital Steel vouchers and the other distribution is based on samples drawn from Efficiency Steel vouchers.

As the next illustrative step we select at random from each of the two distributions one of its sample means. We pair these two sample means and establish their difference, D . This process of random selection and pairing is continued and in this way we get a very large number of such differences. These differences are the raw material from which we form a sampling distribution. (See Illustration 21.14.) This distribution approximates a normal

SAMPLING DISTRIBUTION # 1

SAMPLING DISTRIBUTION # 2

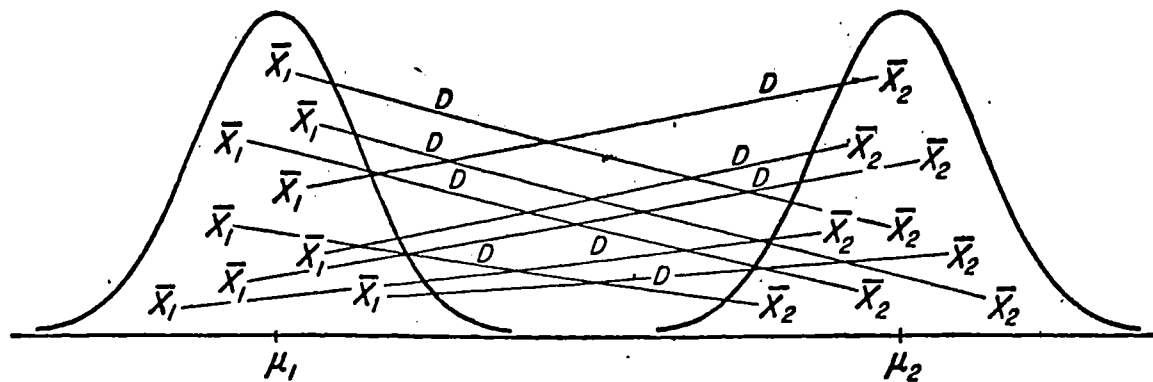


Illustration 21.13. Differences between Means.

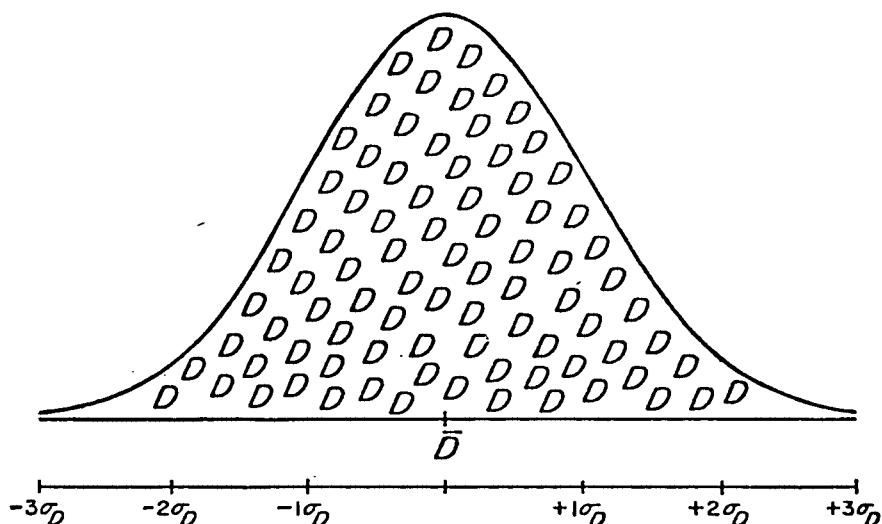


Illustration 21.14. Sampling Distribution Based upon Differences between Means.

distribution. The mean of this sampling distribution of differences between means is the *true difference* between the universe means, symbolized by $\bar{D}$.

A simple illustration makes clear that *the mean of the differences* between paired items of two series, $\bar{D}$, is equal to *the difference between the means* of the two series, $\mu_1 - \mu_2$.

$\bar{X}_1$	$\bar{X}_2$	D
11	5	6
7	5	2
4	3	1
10	7	3
$\mu_1 = \frac{32}{4} = 8$	$\mu_2 = \frac{20}{4} = 5$	$\bar{D} = \frac{12}{4} = 3$

$$\mu_1 - \mu_2 = \bar{D} = 3$$

Of course, the means of the two sampling distributions may be equal and their difference zero. If $\mu_1 - \mu_2 = 0$, then the mean of

differences of sample means, $\bar{D}$, also is zero. A sampling distribution of differences between means with a mean of zero is used as a type which, as we shall see on page 427, is of great importance in statistical inference.

The variation in the sampling distribution of differences between means is measured by the standard error of differences between means. Its computation involves the standard errors of the mean of each of the two sampling distributions, such as are found in Illustration 21.13.

The formula for the standard error of differences between means of two independent distributions is

$$\sigma_D = \sqrt{\sigma_{x_1}^2 + \sigma_{x_2}^2},$$

where σ_D = standard error of differences between means,

σ_{x_1} = standard error of the mean for distribution #1, and

σ_{x_2} = standard error of the mean for distribution #2.

As the plus sign in the formula shows, the standard error of the difference must always be larger than each of the standard errors of the two original sampling distributions from which the differences were established.

In like fashion, the formula for the standard error of the difference between proportions of two independent distributions is based on the standard errors of the proportion of each sampling distribution involved. Thus

$$\sigma_{D_p} = \sqrt{\sigma_{p_1}^2 + \sigma_{p_2}^2},$$

where σ_{D_p} = standard error of differences between proportions,

σ_{p_1} = standard error of the proportion for distribution #1, and

σ_{p_2} = standard error of the proportion for distribution #2.

In a study of expense vouchers, distribution #1 might be the sampling distribution formed by proportions obtained from samples of Capital Steel vouchers and distribution #2 by proportions obtained from samples of Efficiency Steel vouchers.

Standard Error and Sample Size

The relation between sample size and standard error was first discussed on page 396. Since n appears in the denominator of all standard-error formulas,* larger sample size leads to smaller sampling variation or, in other words, greater precision of sampling results. Hence, as we would expect, a larger sample is more reliable than a small sample comparably chosen.

Thus in general a small sample will have a *larger* standard error compared with a large sample of the same characteristics. There is, moreover, a specific situation in the relation between standard error and sample size when we are dealing with what is technically called a "small sample." What is this specific situation? It has been found that the sampling distribution made up of means obtained from "small samples" is *not* a normal distribution. There are different sampling distributions of means for each different small-sample size; examples would be the distribution types for means of random samples of 4 expense vouchers or 10 expense vouchers. These sampling distributions form *symmetrical but not normal* distributions. They differ from normal distributions in being platykurtic† and more widely

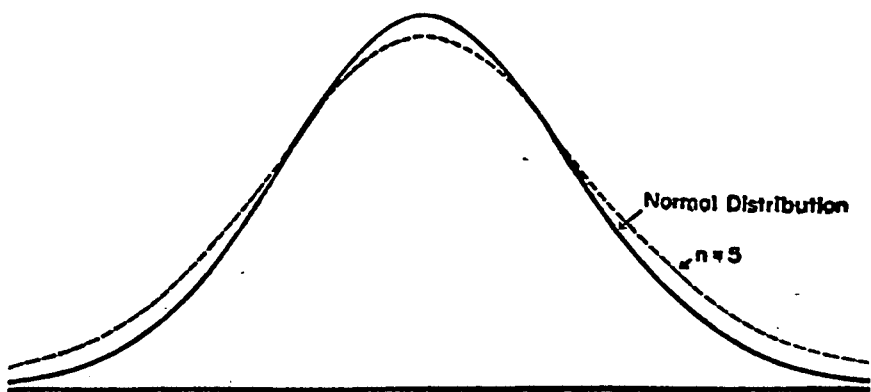


Illustration 21.15. Sampling Distribution Based upon Means of Small Samples ($n = 5$) Compared with Normal Distribution.

* In the formulas for the standard error of a difference, n is inherent.

† It will be recalled that normal distributions are mesokurtic (see page 208)

dispersed. (See Illustration 21.15). The larger the size of the small samples whose means make up the sampling distribution, the more closely this distribution approximates a normal distribution. It has been found that if the sample size exceeds 30, we may safely use the normal distribution as a description of the sampling distribution of the mean.

On pages 398–402, we saw that a range of two standard errors on either side of the true mean includes about 95% of all sample means making up a normal sampling distribution. In a sampling distribution based on means of small samples, the spread is wider and the distribution flatter. (See again Illustration 21.15). Therefore it will take somewhat more than two standard errors to delimit an area including about 95% of all sample means in such a distribution. For example, 2.2 standard errors or 2.6 standard errors may be needed, depending on the size of the “small sample.” Thus *a larger number of standard errors* is needed to include the same percentage of items in a sampling distribution based on means of small samples than in a comparable situation with regard to large samples.

Standard Error and Universe Size

The standard-error formulas are based upon the concept of an infinite universe. When we deal with finite universes, we can create the condition of infinite universes by “sampling with replacement.” In sampling with replacement each item that has been drawn during the sampling process is replaced and thus given a chance of being drawn again as the process continues.

In a finite universe the standard-error formulas are for all practical purposes still valid unless the sample constitutes a large portion of the universe. There is no general agreement on what constitutes a large portion of a universe but certainly one-fifth or more must be considered a large portion. In such cases a correction factor has to be used.*

* See Morris H. Hansen, William N. Hurwitz, William G. Madow, *Sample Survey Methods and Theory*, vol. 1, pp. 122–124.

Standard Error and Sample Size

The relation between sample size and standard error was first discussed on page 396. Since n appears in the denominator of all standard-error formulas,* larger sample size leads to smaller sampling variation or, in other words, greater precision of sampling results. Hence, as we would expect, a larger sample is more reliable than a small sample comparably chosen.

Thus in general a small sample will have a *larger* standard error compared with a large sample of the same characteristics. There is, moreover, a specific situation in the relation between standard error and sample size when we are dealing with what is technically called a "small sample." What is this specific situation? It has been found that the sampling distribution made up of means obtained from "small samples" is *not* a normal distribution. There are different sampling distributions of means for each different small-sample size; examples would be the distribution types for means of random samples of 4 expense vouchers or 10 expense vouchers. These sampling distributions form *symmetrical but not normal* distributions. They differ from normal distributions in being platykurtic† and more widely

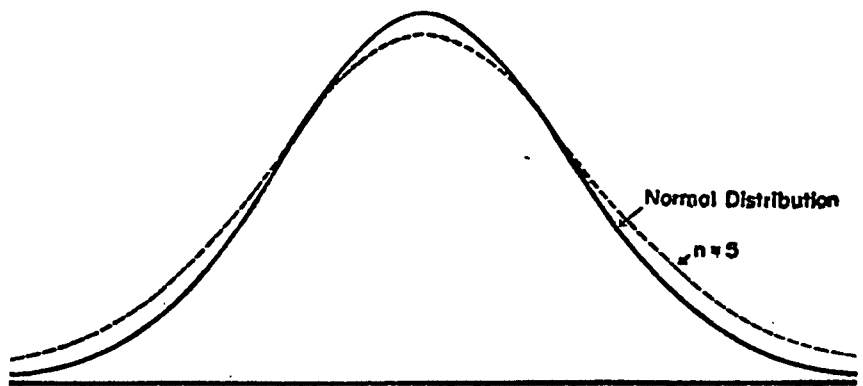


Illustration 21.15. Sampling Distribution Based upon Means of Small Samples ($n = 5$) Compared with Normal Distribution.

* In the formulas for the standard error of a difference, n is inherent.

† It will be recalled that normal distributions are mesokurtic (see page 208)

dispersed. (See Illustration 21.15). The larger the size of the small samples whose means make up the sampling distribution, the more closely this distribution approximates a normal distribution. It has been found that if the sample size exceeds 30, we may safely use the normal distribution as a description of the sampling distribution of the mean.

On pages 398-402, we saw that a range of two standard errors on either side of the true mean includes about 95% of all sample means making up a normal sampling distribution. In a sampling distribution based on means of small samples, the spread is wider and the distribution flatter. (See again Illustration 21.15). Therefore it will take somewhat more than two standard errors to delimit an area including about 95% of all sample means in such a distribution. For example, 2.2 standard errors or 2.6 standard errors may be needed, depending on the size of the "small sample." Thus *a larger number of standard errors* is needed to include the same percentage of items in a sampling distribution based on means of small samples than in a comparable situation with regard to large samples.

Standard Error and Universe Size

The standard-error formulas are based upon the concept of an infinite universe. When we deal with finite universes, we can create the condition of infinite universes by "sampling with replacement." In sampling with replacement each item that has been drawn during the sampling process is replaced and thus given a chance of being drawn again as the process continues.

In a finite universe the standard-error formulas are for all practical purposes still valid unless the sample constitutes a large portion of the universe. There is no general agreement on what constitutes a large portion of a universe but certainly one-fifth or more must be considered a large portion. In such cases a correction factor has to be used.*

* See Morris H. Hansen, William N. Hurwitz, William G. Madow, *Sample Survey Methods and Theory*, vol. 1, pp. 122-124.

Summary

1. The part of a universe selected for investigation is called a sample.
2. Statistical induction involves generalization from the sample to the universe and a statement of the precision of the generalization.
3. Sampling offers advantages over complete enumeration, especially in cost and time, and sometimes even in quality of results.
4. Two main areas of problems in inductive statistics are estimation and statistical significance of a difference.
5. Basic to an understanding of sampling theory are the three distributions: the universe distribution, the sample distribution, and the sampling distribution.
6. The standard error is a fundamental concept in sampling theory and practice.

CHAPTER 22

Estimation and Significance

ESTIMATION

We are now ready to apply to specific problems the rudiments of sampling theory discussed in the preceding chapter. This chapter will be devoted to two important aspects of statistical inference: estimation and statistical significance.

The Concept of Estimation

What does estimation here involve? We draw one sample, observe and measure characteristics of concern, and make inferences about the characteristics in the universe from which the sample came.

Thus in the process of estimation we draw one single sample, measure the mean or median value, the total of the values, the standard deviation of the values, the proportion of items having a certain attribute, or whatever other characteristic or characteristics concern us. From these measurements of one sample, the statistics, we *estimate* the corresponding measurements of the universe, the parameters.

Estimation of the Mean

Let us first take the mean as an illustration. Let us assume we have drawn a random sample of 225 expense vouchers from the

universe of 32,000 expense vouchers of the Capital Steel Corporation written during 1956. The mean value $\bar{X}$ of the 225 vouchers in the sample turns out to be \$25 and the standard deviation is \$7.50.

On page 395 we saw that the sample mean may be used as an estimate of the universe mean, but that we cannot expect to hit the universe mean exactly through the sample mean. We know that if we were to draw many random samples of the same size, the sample mean value which occurred most frequently would be the one at the center of the sampling distribution and would equal μ , the universe mean. Therefore, the sample mean most likely to occur if we draw one single sample is the same as the universe mean. Based on this greater probability of occurrence, we might assume that our sample mean is the best single representative of the universe mean. In the example above described we would thus assume that \$25 is the mean of the universe. In such an assumption we would be making a *point estimate*.

We can do better than this point estimate of the universe mean or true mean by making use of the concept of sampling distribution. Although the sample mean of \$25 might be the one which in repeated sampling would occur most frequently, there is still a very large risk involved in categorically inferring that this value applies to the universe. All we can be sure of is that this sample mean of \$25 lies somewhere within the sampling distribution of means. And we can establish the standard error of this sampling distribution from what we have learned in the preceding chapter:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{7.50}{\sqrt{225}} = \frac{7.50}{15} = 0.50.$$

We are almost certain that the sample mean falls somewhere within a range of $3\sigma_{\bar{x}}$ laid off on each side of the unknown true mean. Hence the statement that our sample mean of \$25 lies somewhere within the sampling distribution of the means can now be made more specific. Since for all practical purposes the

greatest distance of the sample mean is $3\sigma_{\bar{x}}$ away from the true mean, the true mean may be expected to be no further away than $3\sigma_{\bar{x}}$ from the sample mean. Thus we are almost certain that a range of plus and minus three standard errors attached to our sample mean, that is, $\bar{X} \pm 3\sigma_{\bar{x}}$, will include the unknown true mean. This statement holds true no matter where our one sample mean locates in the sampling distribution. Illustration 22.1 depicts the relative location of some sample mean, knowledge of the sampling distribution being assumed.

Thus, in our example of expense vouchers, we are almost certain that a range of $3 \times \$0.50$ added to and subtracted from \$25 will encompass the true mean. In other words, the true mean of expense vouchers will almost certainly lie in this instance between \$23.50 and \$26.50. We have here established an interval of \$3 within which, we are almost completely confident, the true mean lies. This interval of \$3 is called a *confidence interval*, the values of \$23.50 and \$26.50 are called *confidence limits*, and this method of estimation leads to what is called an *interval estimate*.

Accordingly we do not designate one value, \$25, as the true mean or universe mean, but we designate the interval between \$23.50 and \$26.50 as the range which most likely encompasses the true mean. This method of interval estimation is an invaluable advantage derived from sampling theory. Whereas we are unsure of how accurate the point estimate of \$25 for the true mean is, we can state with a known degree of confidence, namely, "almost certainty," that the true mean lies within the confidence interval we have established.

An interval of \$3 in the estimate of the true mean voucher value in our illustration may be too large for some purposes. We may find it more useful to estimate within a narrower interval. If we do so, however, we cannot expect to have as large a degree of confidence in our estimate. If we attach $2\sigma_{\bar{x}}$ to the sample mean on each side, we will have a smaller interval within which the true mean may lie. If unluckily we attach this interval to a rare sample—rare here meaning a sample among the extreme 5%—then the true mean will not lie within this range. Since

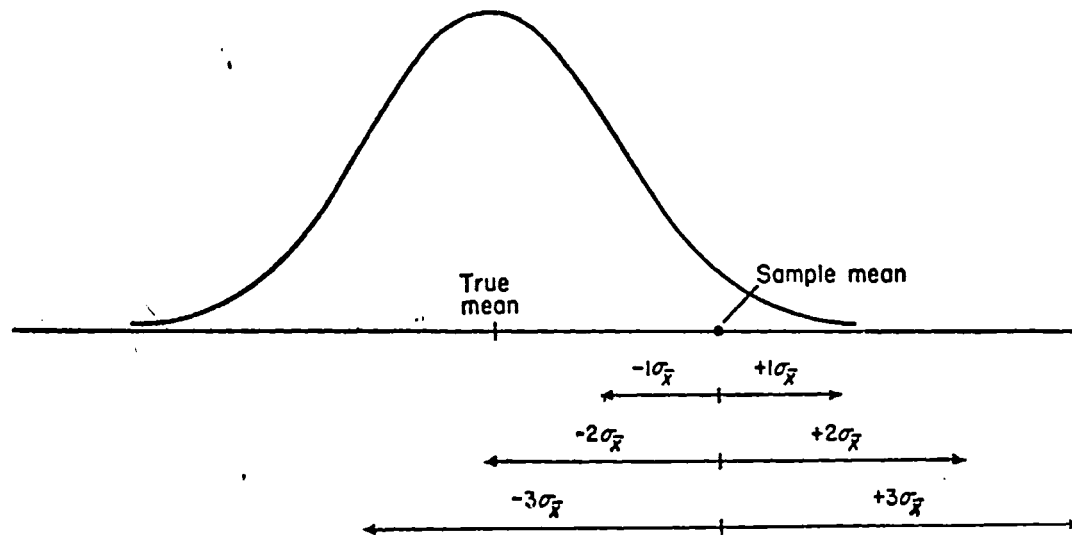


Illustration 22.1. Relation of the Sample Mean to the True Mean in a Sampling Distribution.

only 95% of sample means are not rare, we can only be 95% confident that the interval set up around our sample mean encompasses the true mean. Thus we may say we are 95% confident that an interval of $2\sigma_{\bar{x}}$ * around a sample mean includes the true mean. We may state that we are "reasonably certain" that a range of $\bar{X} \pm 2\sigma_{\bar{x}}$ (in our case, $2 \times \$0.50$ added to and subtracted from \$25, that is, the interval between \$24 and \$26) includes the true mean or universe mean.

Analogously we may choose a confidence interval of $\pm 1\sigma_{\bar{x}}$ or one of any other size. What interval we choose depends upon the individual instance. Thus confidence intervals of ± 1 standard error give us a smaller degree of confidence in the estimate than do intervals of ± 3 standard errors, since we are less certain to have included the true mean within the smaller interval. In practical work a balance must be struck between the desirability of small intervals and the risk involved in using them. In a comparable situation (which does not involve sampling), an estimate that temperatures will be in the thirties involves less risk for the estimator than an estimate that they will be in the *low* thirties; but the latter estimate may be more useful.

The importance of sampling theory in making estimates should now be clear. With one single sample as a basis, we are enabled to stipulate the interval within which the true mean lies and the confidence we can place in such a stipulation.

Estimation of Other Measures

From one sample of expense vouchers we may wish to obtain an interval estimate of the true median of all voucher values, of the true standard deviation of all voucher values, of the true total value, of the true proportion of the vouchers written by executives, or of any other characteristics in the universe. The procedure for setting up a confidence interval for each of these measures follows that already discussed for the mean. First, the particular statistic is established and then a range formed by

* Actually the precise figure is $1.96\sigma_{\bar{x}}$; see the Table of Areas under the Normal Curve, page 213. A range of $\bar{X} \pm 2\sigma_{\bar{x}}$ would include not 95% but 95.45%.

standard errors is attached in order to establish the confidence interval. Finally, we state, with a known degree of confidence, that the parameter is included in this interval.

Formulas for standard errors of several statistics have been given on pages 402-409. Let us illustrate their use in obtaining *interval estimates* of (a) the true median, (b) the true standard deviation, (c) the true proportion, and (d) the true total.

TRUE MEDIAN. In our sample of 225 expense vouchers let us assume the median to be \$24. We already know that $\sigma_{\text{Med.}} = 1.25 \sigma_x$. Since $\sigma_x = \$0.50$ in this example, then $\sigma_{\text{Med.}} = \$0.63$. Hence we may say with 68% confidence that the true median lies somewhere in the range of $\$24 \pm \$0.63$, or between \$23.37 and \$24.63.

TRUE STANDARD DEVIATION. If we wish to obtain an interval estimate of the variation within the universe of expense vouchers, we do so by using the sample standard deviation and its standard error. We know* that $\sigma_s = 0.71 \sigma_x$. Since $\sigma_x = \$0.50$ in this example, then $\sigma_s = \$0.355$. If we wish an estimate with 95% confidence, we attach $2 \sigma_s$ to each side of s . We know that s is \$7.50. Therefore we may state with 95% confidence that the true standard deviation lies somewhere in the range of $\$7.50 \pm 2(\$0.355)$, or between \$6.79 and \$8.21.

TRUE PROPORTION. Let us assume we wish to estimate the true proportion of executive expense vouchers. In our sample we have found that $p = 20\%$. We know that $\sigma_p = \sqrt{pq/n}$ or in this case $\sigma_p = \sqrt{\frac{0.20 \times 0.80}{225}} = \frac{0.40}{15} = 0.027$ or 2.7%. If we wish to estimate the true proportion with practical certainty,† we attach $3\sigma_p$ to each side of p . That is, we may here state that the true proportion lies in the interval $20\% \pm 3(2.7\%)$, or between 11.9% and 28.1%.

TRUE TOTAL. Suppose we wish to make a preliminary estimate of the total value of the 32,000 expense vouchers in the universe. We know that the number of items in a series multiplied by

* See the discussion on pages 402 and 403.

† More precisely, we are making this statement with 99.7% confidence.

their mean equals the total of the items, or $N\bar{X} = \Sigma X$ (see Chapter 10, p. 142). We know $N = 32,000$. We may use the sample mean of \$25 as $\bar{X}$. Thus we obtain an estimate of ΣX , the universe total, by multiplying 32,000 by \$25; the result is \$800,000.

What confidence interval can we attach to \$800,000? We know that the standard error of a total, σ_t , is equal to $N\sigma_{\bar{X}}$. Since in this case, $\sigma_{\bar{X}} = \$0.50$, then $\sigma_t = \$0.50 \times 32,000$ or \$16,000. Thus, if we attach $3\sigma_t$ to \$800,000, that is, a range of \$752,000 to \$848,000, we can say that we are practically certain that the true total value of expense vouchers lies within this range.

Estimation and Sample Size

The application of confidence intervals has up to this point been restricted to large samples. The problem is somewhat different in small samples of 30 items or less.

We saw on page 411 that a larger number of standard errors is needed to include a given percentage of sample means in a sampling distribution based on small samples than is needed to include such a percentage in a sampling distribution based on large samples. Thus with respect to small samples a 95% confidence interval must be based on more than ± 2 standard errors; the size of the small sample determines how much more. A table has been worked out which shows the number of standard errors to be used to establish confidence intervals for different sizes of small samples. An example of the kind of information found in this table follows:

<i>Number of Items in Small Sample, Minus 1</i>	<i>Standard Errors for 95% Confidence Interval</i>
1	12.71
5	2.57
10	2.23
20	2.09
30	2.04
...	...
Large sample	1.96*

* Instead of 1.96 standard errors we use 2 standard errors as an approximation.

A complete table gives more detail by listing all sample sizes up to 30 as well as standard errors for degrees of confidence other than 95%, for example 99%. Such a table is called a *t*-table. The symbol *n* is used in a full table to designate what we here head "the number of items in the small sample, minus 1." *

Let us illustrate the use of the *t*-table by considering the situation of a sample consisting of 6 expense vouchers. Let us assume that in this small sample $\bar{X} = \$28.00$ and $s = \$7.35$. Hence

$$\sigma_{\bar{X}} = \frac{7.35}{\sqrt{6}} = \frac{7.35}{2.45} = \$3.00.$$

To make an interval estimate of the true mean with 95% confidence we find from the *t*-table that for this sample we must attach ± 2.57 standard errors. Since $\sigma_{\bar{X}} = \$3.00$, we attach 2.57 times \$3.00 or \$7.71 on both sides of the mean. Thus we are 95% confident that the true mean estimated from this sample is included in the interval of $\$28.00 \pm \$7.71$, or between \$20.29 and \$35.71.

STATISTICAL SIGNIFICANCE

The Concept of Statistical Significance

Let us now turn our attention to the second important aspect of statistical inference, namely, the testing of statistical significance.

What does testing statistical significance mean? We come upon differences of some measurement and wish to establish whether differences found on the sample level also hold on the universe level—that is, whether the differences are real or whether they may be due to variation arising from the sampling process. A *real* difference means one inherent in the very nature of the data as opposed to a *chance* difference which has been brought about by the process of sampling.

* This usage of *n* involves what are called "degrees of freedom." For a discussion, see Helen M. Walker and Joseph Lev, *Statistical Inference*, New York, Henry Holt and Company, 1953, p. 90-91, 135, 145.

In testing statistical significance we find two major types of problems. Examples of these two major types follow.

1. PROBLEMS CONCERNING THE COMPARISON OF TWO SAMPLES. The Capital Steel Company has acquired the Efficiency Steel Company. From the universe of expense vouchers of the Efficiency Steel Company a random sample of 225 vouchers has been drawn and a mean value of \$28 is found for this sample. The corresponding figure for the sample of the Capital Steel Company we know to be \$25. The controller of the Capital Steel Company would like to know whether this difference of \$3 is a *real* difference or whether it is merely due to the fact that the two mean values of \$28 and \$25 are based on samples. If *all* expense vouchers for each of these two companies had been studied, would this same difference have appeared? If this difference of \$3 appeared in the means based on all expense vouchers, it would be a real difference inherent in these two voucher universes.

2. PROBLEMS CONCERNING THE COMPARISON OF A SAMPLE AND A UNIVERSE. In 1954 the mean expense-voucher value was computed from all the 31,000 expense vouchers filed during that year in the Capital Steel Company. The mean value was \$23. This universe mean of \$23 is less than the sample mean of \$25 which was obtained in 1956. We want to establish whether this difference of \$2 is real or whether it may be attributed to chance. If there is likelihood that this difference is real, we call it statistically significant. We would then hold that the mean voucher value has changed over these two years.

The Null Hypothesis

How do we go about establishing whether the differences illustrated above are statistically significant or due to chance? We first formulate a hypothesis about the difference. Customarily this hypothesis states that the true difference between the two values in question is zero. This formulation is called a *null hypothesis*.

Under the null hypothesis we start making our test of statistical significance by saying: "There is no difference." Then we

must arrive at a decision to accept or to reject the null hypothesis. The rejection of the null hypothesis indicates that the differences have statistical significance and the acceptance of the null hypothesis indicates that the differences are due to chance. Since many practical problems aim to establish statistical significance of differences, rejection of the null hypothesis may thus indicate success in a statistical project.

For a null hypothesis to be accepted the observed difference between values need not be zero even if the real difference is zero, since sampling may account for the departure from zero. Thus we accept the null hypothesis even if the observed difference is not zero provided the observed difference is likely to be due to chance. But if the observed difference appears too large to have been brought about by chance, we reject the null hypothesis and conclude that a real difference exists..

How do we decide when a difference may be ascribed to chance and when a difference may be statistically significant? We turn now to a detailed discussion of this question.

Difference between Sample Mean and Universe Mean

The problem of testing the statistical significance of the difference between a sample mean and a known or assumed universe mean is substantially a different form of the problem of estimation discussed earlier in this chapter. We refer again, therefore, to Capital Steel Company's 1954 mean expense-voucher value of \$23 and its 1956 sample mean of \$25 (see page 421). We wish to know whether the difference between this universe mean of 1954 expense vouchers and the mean value of the 1956 random sample of 225 vouchers is statistically significant. According to whether the difference turns out to be a real difference or a difference due to chance, we can say whether or not the 1954 voucher mean may still be appropriate to describe the 1956 situation. How do we go about learning whether the difference is real?

The mean of the sampling distribution is \$23 since its mean is

the same as the universe mean μ ; its standard error $\sigma_{\bar{x}}$ is \$0.50. This standard error of the sampling distribution comes from information yielded by the sample (see page 397). We must now locate the sample mean within the sampling distribution. Depending on its location within the sampling distribution, we may classify a sample mean as (1) common or (2) unusual or rare. (The criteria are explained below.) If the sample mean is common,

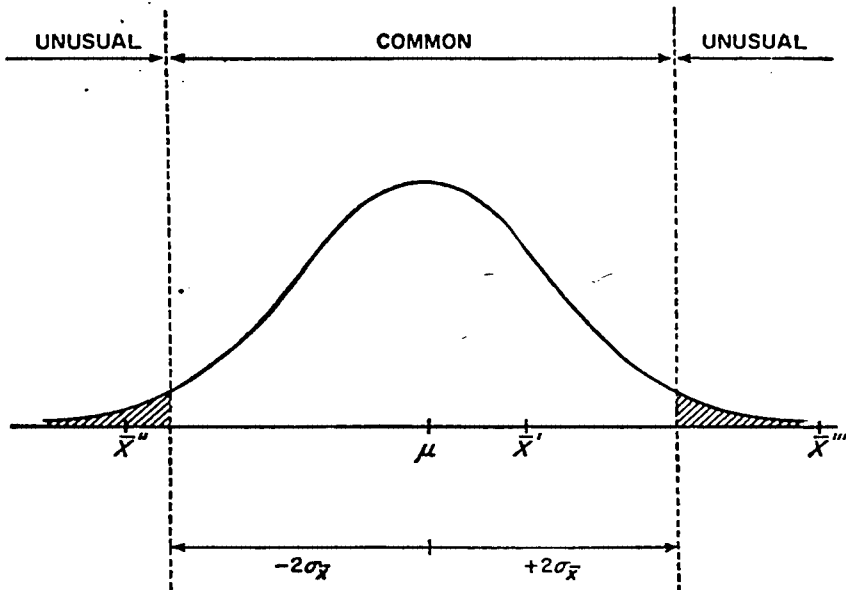


Illustration 22.2. True Mean, Common Sample Mean, and Unusual Sample Means.

it is likely to “come from” the universe with mean μ , and the difference between the sample mean and the universe mean is not statistically significant. Sample mean $\bar{x}'$ in Illustration 22.2 shows this situation. If, however, the sample mean is unusual, as are $\bar{x}''$ and $\bar{x}'''$ in Illustration 22.2, then we conclude that it belongs either to an unusual sample drawn from this universe or to a sample coming from a different universe. The difference between such a sample mean and the universe mean is statistically significant.

What are the criteria of the "unusual" upon which we base a decision concerning statistical significance? There is no general rule that establishes when a sample mean is common and when unusual. In setting up this dividing line the statistician may disqualify a sample mean which really belongs to the sampling distribution or he may call a sample mean common which really belongs to another universe. The prevalent practice is to call rare the extreme 5% of sample means in the sampling distribution; that is, the 2.5% at the low extreme and the 2.5% at the high extreme of the sampling distribution. This practice establishes what is called the 5% level of significance. This break into 5% and 95% in a distribution of sample means corresponds to the 95% degree of confidence in making estimates discussed earlier in this chapter.

In practical work, statisticians sometimes use a 1% significance level, which classifies as rare the $\frac{1}{2}\%$ of sample means at each end of the distribution.

In order to establish whether a given sample mean is common or unusual we must locate it within the sampling distribution. This location is arrived at with the aid of the standard error of the mean.

Let us use the 5% significance level for purposes of illustration. We have already seen that by cutting the sampling distribution at points $\pm 2\sigma_{\bar{x}}$ away from the true mean μ , we separate approximately 5% of the extreme sample means. Thus every sample mean *more than* $2\sigma_{\bar{x}}$ away from the true mean is considered *unusual* and its difference from the true mean is *statistically significant*.

In numerical terms we find how many standard errors away from the true mean a given sample mean is. We do this by expressing the difference between the sample mean and the true mean in terms of the standard error. Thus in the case of the expense vouchers where the universe mean μ is \$23, the sample mean $\bar{X}$ is \$25 and the standard error is \$0.50, the calculation is

$$\frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{25 - 23}{0.5} = 4.$$

This result shows the sample mean to be 4 standard errors away from the universe mean. Since the 5% level of significance has been used which involves a distance plus or minus $2\sigma_{\bar{x}}$, this difference of $4\sigma_{\bar{x}}$ from the universe mean is statistically significant.

Difference between Sample Proportion and Universe Proportion

In similar fashion, if the universe percentage of vouchers written by executives is found to be 24% (that is, $P = 24\%$), and the sample percentage is 20% (that is, $p = 20\%$), with the standard error σ_p being 2.8%,* then

$$\frac{p - P}{\sigma_p} = \frac{20 - 24}{2.8} = -1.4.$$

This distance of 1.4 standard errors of a sample proportion from the universe proportion is common and probably due to chance. Consequently the difference between the universe proportion and the sample proportion is not statistically significant.

Difference between Means of Two Samples

In testing the statistical significance of a difference between two sample means, we first envisage a sampling distribution of differences between means as described on pages 406-409. Then we locate the observed difference in this sampling distribution. To establish the significance of a difference we set up a criterion of the unusual.

Let us again use the 5% significance level, which means that we cut the sampling distribution at ± 2 standard errors away from the mean.† We must now locate the observed difference in the sampling distribution of differences between means. The mean of such a distribution is the *true difference* between the two universe means, $\bar{D}$. In terms of the null hypothesis we

* On page 418, σ_p was found to be 2.7%. Here, since P is known, it is used instead of p in the computation of σ_p .

† If we had chosen a 1% level of significance, we would have had to cut the distribution at ± 2.6 standard errors away, thus classifying the extreme $\frac{1}{2}\%$ of items on each side as unusual.

assume that there is no difference between the two universe means. Thus, the true difference is assumed to be zero, or in other words all possible differences which go to make up this sampling distribution are considered to average zero. See Illustration 22.3.

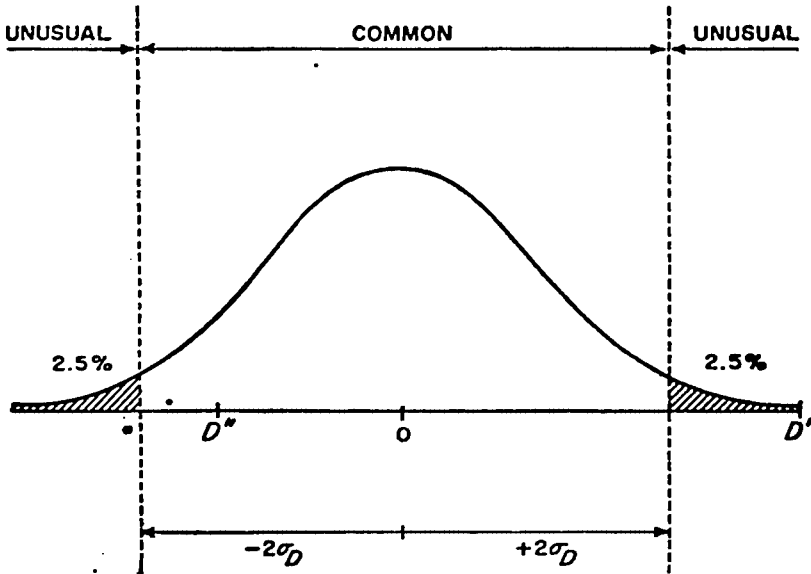


Illustration 22.3. Common and Unusual Differences in a Sampling Distribution of Differences between Means.

The standard error of the sampling distribution of differences between means is σ_D (see page 409). We may classify as unusual or rare any difference between two sample means that deviates by more than $\pm 2\sigma_D$ from the true difference of zero. Such a rare difference between two sample means is not compatible with the null hypothesis that the difference between the two universe means is zero. A rare difference between means such as D' in Illustration 22.3 is unlikely to come from a sampling distribution of differences with a true difference of zero. Thus the two sample means differ significantly. On the other hand, a difference of D'' between two sample means as in Illustration 22.3 does not exhibit statistical significance.

What is the technique for measuring the statistical significance of a difference between means? We know that the sample drawn at random from the expense vouchers of the Efficiency Steel Company shows a mean voucher value of \$28.00 and that the sample for the Capital Steel Company shows a mean of \$25.00. The standard error of the mean $\sigma_{\bar{X}}$ is found to be \$0.80 for Efficiency while the corresponding measure for Capital is $\sigma_{\bar{X}} = \$0.50$. On page 409 we saw that

$$\sigma_D = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2};$$

or in this case,

$$\sqrt{0.64 + 0.25} = \sqrt{0.89} = 0.94.$$

In order to locate the difference between the sample means in the sampling distribution of differences, we express this difference in terms of the standard error of the difference. In symbols:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sigma_D} = \frac{28 - 25}{0.94} = \frac{3}{0.94} = 3.2.$$

Thus the difference between the sample means of \$28.00 and \$25.00 is unusual—more than $2\sigma_D$ away from the hypothesized true difference of zero. We therefore reject the null hypothesis and consider the difference statistically significant.

Difference between Two Sample Proportions

Now let us illustrate how we test the statistical significance of a difference between proportions. If in Efficiency Steel the proportion of expense vouchers written by executives $p_1 = 17\%$ and $\sigma_{p_1} = 2.5\%$ and in Capital Steel $p_2 = 20\%$ and $\sigma_{p_2} = 2.7\%$, then, as on page 409,

$$\sigma_{D_p} = \sqrt{\sigma_{p_1}^2 + \sigma_{p_2}^2}$$

or here

$$\sigma_{D_p} = \sqrt{(2.5)^2 + (2.7)^2} = 3.7.$$

The difference between the sample percentages, expressed in terms of the standard error, is

$$\frac{p_1 - p_2}{\sigma_{D_p}} = \frac{17 - 20}{3.7} = \frac{-3}{3.7} = -0.81.$$

Thus the difference between the sample percentages of 17% and 20% is not unusual at our significance level (that is, is not more than $2\sigma_{D_p}$ away from the hypothesized true difference of zero). We therefore accept the null hypothesis and consider the difference between these proportions not statistically significant.

The formulas for testing significance are applicable only if the samples involved are large, that is, only if they consist of 30 items or more. Otherwise adaptations must be made in the light of the special situation of small samples discussed earlier in this chapter.*

Limitations of Tests of Significance

In testing statistical significance several limitations must be kept in mind.

Statistical tests of significance can supply guides to the discriminating user but they must not be used mechanically; they are raw material from which to make decisions, not decisions in themselves. But this limitation does not detract from the aid they can give when properly interpreted.

There are situations where real differences exist but do not produce evidence that they are statistically significant; this anomaly may occur when the sample is small. On the other hand, the tests may show that a difference has statistical significance even when it is so small that it must be interpreted with great care.

When tests show that a difference has statistical significance they suggest that the observed difference is probably not due to chance. Thus statements are not made with certainty but with

* See also p. 411 in the preceding chapter concerning samples which are a large part of the universe.

a known degree of probability. "Unusual" events do happen once in a while.

Furthermore, though tests can indicate that a difference has statistical significance, they do not tell us *why* the difference exists. But tests of significance show the need for further investigation in order to reach definite answers.

In problems of statistical significance, as in other statistical problems, technique must be combined with good judgment and knowledge of the subject matter.

Summary

1. In statistical estimation a universe measure is estimated from a sample measure.

2. An interval estimate establishes, with a stated degree of confidence, the limits within which the universe measure lies. These limits are called confidence limits and the range between them is called a confidence interval.

3. Small samples require special treatment.

4. The testing of statistical significance aims at distinguishing between a real difference and a chance difference.

5. The null hypothesis is used in testing statistical significance.

Sampling Practice

Up to this point in the discussion of sampling we have concentrated on the rudiments of sampling theory as the foundation for statistical inference. But we have not yet paid sufficient attention to what in practice constitutes an appropriate sample. There are decisions to be made concerning the kind of sample, the design of the sample, the drawing of the sample, the size of the sample.

Many ingredients are needed for a good sample. The sampling statistician must combine knowledge of sampling theory with craftsmanship. Craftsmanship includes use of general knowledge of statistics, of the field of application such as marketing or industrial production, as well as imagination, intellectual adaptability, and resourcefulness. Past experience with the data leads to sound judgment concerning the sample needed. Certain decisions on matters of fact are an important part of sampling. There are questions about lists of sampling units, about the availability of detailed maps, about when to take the sample, and other such questions.

It was only recently—in the late 1940's—that comprehensive books on sampling became available. Today sampling is a main concern in statistics.

Statistical sampling aims to provide universe estimates with confidence intervals that are as narrow as possible. In other words, the goal is *small standard errors*. This aim must, however,

be viewed in the light of practical considerations, of which one is cost. The greater the sampling precision the greater the cost. Great precision may be expensive. But great precision may not be necessary. Thus a balance must be sought between cost and a desirable degree of sampling precision.

Another qualification of this aim of small standard errors becomes apparent when we view the sampling error as just one component of the total error in a statistical inquiry. The *total error* includes *response errors* in sample surveys, *measurement errors* in experiments based on samples, and *processing errors* which occur during the processing of sample data; these are some of its major components.

It may be desirable in some instances to accept a larger sampling error, that is, larger standard errors, in return for a reduction of other error components. For example, we may reduce the sample size so that it is commensurate with the number of *trained* interviewers available for the survey. In such a case the sampling error would be larger but the interviewer error would be substantially smaller.

RANDOM SAMPLING

All our discussions on samples up to this point have referred to random samples.* Random sampling is also called *probability* sampling. What has been said up to now about sampling theory and its application is valid only for probability sampling. But random samples are not the only type. We distinguish two major types: (1) random samples, and (2) purposive samples. The latter type will be treated in some detail later in this chapter.

Random Selection

Random samples are characterized by the way in which they are selected. "Random" is not used here in the sense of "hap-

* See page 385.

hazard" or "hit-or-miss." A random sample is a sample where each item in the universe has an equal or known opportunity of being selected—providing chance, and only chance, determines which items are to be included in the sample chosen. Selection by chance enables us to make use of the concept of probability, which is the foundation upon which sampling theory rests.

Suppose we are making a study in the city of Houston, Texas, and wish to take a sample of all residents of this city. We shall not arrive at a random sample by using the Houston telephone directory and selecting names "at random." Why not? First, when we use the telephone directory we do not give every item in the universe, in this case every resident of Houston, an opportunity of being included since not every resident of Houston has a telephone and since some telephones are "private numbers" and not listed. Secondly, we may pick names from the telephone directory by some sort of selection, wittingly or unwittingly, taking only long names or short names or names associated with a national origin close to our interest. This selective factor is *bias*. The human mind unaided or unguided has a tendency away from true randomness; thus certain types or groups of items in the universe have a better opportunity of being selected.

It is not easy in practice to insure true randomness. To obtain it the statistician makes use of certain aids. If the universe is not too large, the information referring to each item may be listed on an index card. These cards are referred to as a "card universe." By shuffling these cards we open the way to a random selection. Chips numbered to match numbers given each item may serve the same purpose; the random sample is drawn from the chips after they have been thoroughly mixed.

Wide use is made of so-called tables of random numbers. These tables consist of many columns of numbers which have been subjected to practical tests of randomness. We accept the sequence of these numbers as truly random. A table of random numbers and an explanation of their use are shown in an Appendix to this chapter.

Restricted Random Samples

"The sample design is the blueprint for the method of selection of the sample."* Previous discussions have referred to the design of what are called *unrestricted random samples*. These discussions are basic to understanding of sampling. There are, however, what may be called *restricted random samples* which are of great practical value. These restricted random samples make use of various selection methods, some of which are discussed below. In general the intention is to increase efficiency (that is, to achieve greater precision of sampling results per dollar). A restricted random sample may aim either at greater precision (that is, smaller standard errors), or at smaller cost compared with an unrestricted random sample of the same size. In designing a restricted random sample the statistician makes use of supplementary information concerning the universe, for example, relevant information from the Bureau of the Census.

An additional problem in practical sampling is time. Results are frequently needed within a short time. The sample design may reduce the time involved in obtaining results although the designing of a sample itself may be time-consuming in some instances.

Finally, there are situations where unrestricted random sampling is not feasible and a restricted random sample must be used. Such situations are often met when we deal with sampling units like persons or stores and good lists of their addresses are unavailable.

It is important to note that the method for computation of the standard errors for the various types of restricted random samples differs from the previously discussed method. But the formulas for standard errors in unrestricted random sampling are the basic models.† Three of the most important selection

* Lester R. Frankel and Sol Dutka, "How to Get a True Sample of Households," *Printers' Ink*, July 15, 1955, p. 31.

† For details see Morris H. Hansen, William N. Hurwitz, and William G. Madow, *Sample Survey Methods and Theory*, New York, John Wiley and Sons, Inc., 1953, volume 1.

methods in restricted random sampling lead to: (1) stratified sampling; (2) cluster sampling; (3) systematic sampling.

Stratification

Stratification means dividing an entire universe into groups. These groups or subuniverses are called *strata*. The purpose of stratification is to increase the efficiency of sampling by dividing a heterogeneous universe in such a way that (1) there is as great *homogeneity* as possible *within* each stratum and (2) as marked a *difference* as possible *between* the strata.

To illustrate this division let us return to the study of expense-voucher values. We might sample more efficiently if we could set up strata that are as homogeneous as possible with regard to the characteristic under investigation, that is, expense-voucher value. From experience, we know there is a relation between value of voucher and salary level of employee. Therefore we set up strata of vouchers originated by high-, medium-, and low-salaried employees. We expect some degree of uniformity in voucher values originated by the high-salaried, the medium-salaried, and the low-salaried. Thus we achieve homogeneity within each of the three strata and substantial difference of the mean voucher values among the three strata. We would certainly not classify the universe of expense vouchers into those written by tall and by short persons. The stratification should not be based on a characteristic such as height of employees which has no relationship to the characteristic under investigation, namely, voucher value.

Often strata are subdivided into substrata which in turn may be subdivided into still smaller groups. For example, the strata set up with respect to high-, medium-, and low-salaried employees may be subdivided according to sex of employees, thus giving six substrata of vouchers, originated by male high-salaried, female high-salaried, etc.

The individual problem under investigation points to characteristics that should be used for stratification. In practice, geo-

graphical, sociological, and economic characteristics are often used for the grouping of the universe.

What representation shall each stratum or substratum have in the sample? Should each stratum yield a number of sampling units proportionate to its size, or an equal number of sampling units? Or should some other criterion be used? Which method to use in representing strata must be decided in accordance with the problem to be solved.

If 8% of all vouchers originate from the high-salaried stratum, 8% of all sample vouchers may be drawn from this high-salaried stratum. Similarly, each stratum can be represented according to its size in the universe. This method of representing strata in the sample is called *proportionate sampling*.

Another method consists of representing each stratum equally, regardless of its over-all size in the universe. Thus one-third of the sample would be drawn from the high-salaried stratum, one-third from the medium-salaried, and one-third from the low-salaried.

In some instances it may be excessively expensive to investigate sampling units of one particular stratum, for example, persons living in places with population of less than 100. A small representation of such a stratum is nevertheless desirable.

Both equal and small representation are illustrations of what is called *disproportionate sampling*.

A special and rather important method of representing strata is the method of *optimum allocation*. Each stratum is represented in the sample according to its size and its variability. It is clear from our previous discussion that a larger sample is needed to represent a *universe* with large variation than one with small variation. In optimum allocation, each *subuniverse* or stratum is represented in the sample according to its dispersion as well as its size.

After the strata have been set up and the method of representing them decided upon, a *random* sample is drawn from each

stratum in the same way as in unrestricted random sampling. Separate estimates are obtained from each stratum and then these estimates are appropriately combined.

ADVANTAGES AND LIMITATIONS OF STRATIFIED SAMPLING. The method of stratified sampling has, in many instances, considerable advantages over unrestricted random sampling. Very often sampling precision is greater in a stratified sample than in an unrestricted sample of equal size. Hence, often the standard error of the stratified sample is substantially smaller than that of a comparable unrestricted random sample.

Greater precision of estimates from stratified samples arises in part from the use of supplementary information. General information about the universe, the experience and skill of the statistician, and sound judgment come into play in setting up strata. Formulas alone do not provide solutions for the problem of the best method of stratifying. This subjective aspect in the process of stratifying does not mean that the sampling units are selected subjectively. The sampling units must be selected by chance.

Stratified sampling is usually of great advantage when the universe distribution is skewed. Where such skewness exists, a small number of very important items may be located in the tail of the distribution. For example, in a study of incomes, the universe distribution will most likely include at the upper extreme a small number of very high incomes. Unrestricted random sampling may miss this entire small group whereas stratified random sampling will ensure its representation by setting up a stratum of high incomes.

Furthermore, stratified sampling yields more than just an estimate of universe characteristics; it yields separate estimates for each subuniverse and makes possible comparison among them.

There are, however, limitations attached to stratified sampling. For one, precision of a stratified sample is not *always* greater than that of an unrestricted random sample of equal size. For another, the supplementary information needed to set up strata

effectively is sometimes unavailable and often expensive to procure.

Cluster Sampling

In cluster sampling the universe is, as in stratified sampling, divided into groups. In cluster sampling, however, these groups or clusters become the sampling units. To avoid confusion of terms, a group which plays the role of a sampling unit will here be called a *primary sampling unit* (psu) to distinguish it from the single-item sampling unit mentioned previously in Chapter 21. The single-item sampling unit may be designated as the *elementary sampling unit*, to distinguish it from the primary sampling unit.

Thus in the study of expense-voucher values, the individual voucher is the elementary sampling unit. At the time of the illustrative sample survey, all expense vouchers of the company are in file drawers of equal size, stored in different places. Each drawer is a cluster of vouchers. We could make the file drawers our primary sampling units. As the first step in cluster sampling we would select a random sample of file drawers, that is, of primary sampling units. Thus we would select a random sample consisting of clusters.

As the second step, we can either investigate each and every elementary sampling unit in each selected cluster, or we can draw a random sample of elementary sampling units from each selected cluster. That is, we can study all expense vouchers in each of the selected file drawers or we can draw a random sample of vouchers from each selected drawer.

The aim of taking groups in cluster sampling is opposite to the aim of selecting groups as strata. It is desirable to have (1) as great a *heterogeneity* as possible *within* clusters and (2) as *small a difference as possible between* the clusters. Thus if the voucher values within each drawer are very much alike and the drawers differ considerably from each other, our cluster sample of vouchers will not give good representation.

Often in practical work clusters are chosen on a time basis.

Thus a cluster may consist of the sales checks referring to a certain food product during three-day periods, or machine parts finished during two-hour periods.

Clusters on a geographical basis are of great practical importance. Suppose we wish to study grocery-store sales in the United States. The elementary sampling unit is the grocery store. We may select clusters of grocery stores on a geographical basis, by county, let us say. This type of cluster sampling is called *area sampling*, which has gained in importance since 1939 when the first area sample was selected in the United States.* Aided by maps, the statistician first draws a random sample of areas (in this case, counties). He then studies all grocery stores—elementary sampling units—in each selected area or he draws a random sample of grocery stores within each selected area.

If further stages are added to the design just discussed, we are involved in what is called multistage sampling. In multistage sampling we first select primary sampling units, then from them we select secondary sampling units. In the secondary sampling units we study all elementary sampling units or else a sample of elementary sampling units. Each sample throughout this process must be a random sample. To illustrate: having selected a random sample of file drawers (primary sampling units) we may, within each selected drawer, choose clusters of ten folders each (secondary sampling units). Finally, all vouchers contained in the ten folders or a random sample of them (elementary sampling units) may be studied.

ADVANTAGES AND LIMITATIONS OF CLUSTER SAMPLING. Cluster sampling may offer several important advantages. As a general rule, a cluster sample is considerably less expensive than an unrestricted random sample of equal size. It can easily be seen from the illustration above that an unrestricted random sample of vouchers kept in different places would have involved the pulling of vouchers from files at widely dispersed locations. A sample design which cuts down on moving about means savings

* Lester R. Frankel and Sol Dutka, "How to Get a True Sample of Households," *Printers' Ink*, July 15, 1955, p. 31.

in time and money. This advantage becomes clearer in situations where travel of interviewers between the locations of elementary sampling units is involved.

Furthermore, the lack of complete, up-to-date lists of elementary sampling units often makes unrestricted random sampling impossible. Thus no unrestricted random sample of feed stores can be drawn if good lists of feed stores are unavailable. It is possible, however, to make a listing of feed stores within the selected primary sampling units, say, counties, and to use this list in the selection of a sample of feed stores. Thus area sampling is sometimes the only solution to practical sampling problems.

There is, however, an important limitation that may attach to cluster sampling. Usually the standard error of a cluster sample is larger than that of an unrestricted random sample of equal size; thus estimates based on a cluster sample are frequently less precise. This limitation arises from the fact that the requirement of heterogeneity of the items within a cluster is rarely fulfilled. Items within a cluster tend toward similarity. Birds of a feather cluster together.

A decision on the use of cluster sampling often requires weighing the gain in economy against the loss in precision.

Systematic Sampling

While stratification and cluster sampling involve the division of the universe into groups, systematic samples involve an ordering of the universe. The ordering may be alphabetical, geographical, or any other.

A systematic sample is selected by including every n th item from the universe, every 100th expense voucher for example. The selection may be done by pulling out the actual sampling units such as expense vouchers or by choosing from lists or files the entries or cards which represent the sampling units; from either the sampling units or their representatives every n th item is selected. Sometimes the statistician need not have details of the universe but can nevertheless lay down definite rules for a systematic selection of sampling units; for example, in sampling dwelling

units a sample theoretically sound can be obtained by the inclusion of some randomly selected dwelling unit and every 10th dwelling unit thereafter along a planned route.

In systematic sampling we achieve greater efficiency if we (1) secure as great a *similarity* as possible among *immediate* "neighbors" in the universe as ordered in terms of the characteristic to be measured, and if we (2) increase the differences between *distant* "neighbors." If we place expense vouchers in order of magnitude and then draw, say, every 100th expense voucher, we have met these conditions.

The selection of a systematic sample involves decisions on (1) a sampling interval and (2) a sampling start.

The choice of sampling interval depends on the size of the universe and the size of sample to be chosen. If we wish to draw a sample of 320 expense vouchers from a universe of 32,000 vouchers, the sampling interval is 100; that is, every 100th voucher is to be selected.

Thus we may start the sample selection anywhere between the first and the 100th voucher. A random start is always preferable; that is, a start determined by chance. The starting number may be chosen from a table of random numbers. The 3rd item has as much chance to be the start as the 76th item. Thus the 3rd, 103rd, 203rd . . . items have the same chance of being selected as the 76th, 176th, 276th have. Therefore, in systematic sampling each item in the universe has a known chance of being included in the sample; this makes it a probability sample.

ADVANTAGES AND LIMITATIONS OF SYSTEMATIC SAMPLING. The main advantage of systematic sampling is its simplicity. Furthermore, an advantage of stratification can be realized in systematic sampling by an appropriate ordering of the universe. For example, as pointed out above, if we arrange the universe of expense vouchers by magnitude, we secure in effect proportionate representation of all voucher-size classes in the sample.

The precision of an estimate from a systematic sample is frequently greater than from a comparable unrestricted random

sample. Sometimes, however, the standard error of a systematic sample may be larger than that of an unrestricted random sample of equal size.

Systematic sampling may lead us astray if there is a pattern in the universe. For some reason every 50th expense voucher could be one made out by an executive; an unlucky systematic sample that hit these vouchers would thus overrepresent vouchers of executives, which have special characteristics. Similarly, in an inventory check based on systematic sampling, one type of item may be overrepresented because it is in every case stored at the same corner of the shelf or listed on the same line of each page of the inventory record. Or if sales checks of every 7th calendar day form the systematic sample, a pattern may appear that is inherent in this day of the week.

Other Sample Designs

Several other types of sample design are in use in practical work and the reader is referred to special books on sampling for their discussion. (See Appendix VI, Bibliography.)

It should be mentioned, however, that a particular sample may combine aspects of more than one design. For example, stratification may be used in the selection of primary sampling units in cluster sampling and the elementary sampling units within the primary sampling units may be chosen by systematic sampling.

Sample Size

How large should a sample be? This question commands attention in sampling practice. Pat answers of the type: "5% of all items" or "1000 units" are of no use as general rules. Appropriate sample size depends on factors in the individual situation. Sampling theory is of aid in arriving at good estimates of sample size. The standard error here too provides the key.

As we know, the formula for the standard error of the mean $\sigma_{\bar{x}}$ contains the sample size n , as does the formula for the stand-

ard error of the proportion σ_p . Thus n can be found if we can estimate or specify the other elements of the equation. We know the basic formulas

$$\sigma_x = \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \sigma_p = \sqrt{\frac{PQ}{n}}$$

The restatement of these formulas leading to a solution for n gives:

$$\sigma_x^2 = \frac{\sigma^2}{n} \quad \text{or} \quad n = \left(\frac{\sigma}{\sigma_x} \right)^2$$

and

$$\sigma_p^2 = \frac{PQ}{n} \quad \text{or} \quad n = \frac{PQ}{\sigma_p^2}$$

A solution of these two equations for the sample size n involves three steps. The first step requires an estimate of an unknown measure; the second and third steps require a specific decision in each case.

How can we establish appropriate sample size if we wish to estimate the mean?

Step 1: We require the standard deviation of the universe, which is, of course, unknown. A rough estimate of this measure, however, is sufficient for suggesting sample size.

(a) In many instances, experience with similar problems will be a good guide for making this estimate of the standard deviation.

(b) In other instances, an exploratory sample study on a small scale may be conducted in order to arrive at an estimate of σ .

(c) To estimate the standard deviation of the universe, the range of the values in the universe may be estimated and used as a guide. We know (page 195) that in a normal distribution the range is about six times the standard deviation. In practical work, an estimate of somewhere around one-fifth of the estimated range is often used. Suppose we have reason to believe that the range is roughly \$300; that is, the difference between the lowest value in the universe and its highest value is \$300.

One-fifth of this rough estimate is \$60. Therefore we take \$60 as a rough approximation of σ .

Step 2: We must decide upon how precise we want the future sampling estimate to be. Thus we may state that the estimate of the true mean is sufficiently precise if confidence limits of $\pm \$12$ are attached to it. Such an answer might be practicable for this particular problem.

Step 3: We must decide how confident we want to be that the future estimate will be within the specified limits. We may wish to be almost certain or be satisfied with, say, a 95% degree of confidence, that the specified limits will contain the true mean. The degree of confidence decided upon makes it possible to translate the interval decided upon in step 2 into standard errors. If we are to be practically certain that the true mean will lie within the interval of $\pm \$12$ around the sample mean, then the interval of $\pm \$12$ becomes $\pm 3\sigma_{\bar{x}}$. Therefore, $\sigma_{\bar{x}} = \$4$. If, on the other hand, we are willing to settle for a 95% degree of confidence, $\pm \$12$ becomes an interval of approximately * $\pm 2\sigma_{\bar{x}}$ and $\sigma_{\bar{x}} = \$6$.

We are now ready to use the formula for n :

$$n = \left(\frac{\sigma}{\sigma_{\bar{x}}} \right)^2$$

Since $\sigma = \$60$ (Step 1) and since at the level of practical certainty $\sigma_{\bar{x}} = \$4$ (Steps 2 and 3), therefore

$$n = \left(\frac{60}{4} \right)^2 = (15)^2 = 225.$$

If we settle for a 95% degree of confidence that the true mean will be within $\pm \$12$ of the sample estimate, then

$$n = \left(\frac{60}{6} \right)^2 = (10)^2 = 100.$$

Thus in this case the sample size should be somewhere around 225 if we wish to be practically certain that the true mean will

* This holds true only for large samples. The number of standard errors needed in the case of small samples is larger (see page 411).

lie within an interval of $\pm \$12$; but the sample need contain only 100 items if we settle for a 95% degree of confidence that the true mean will lie within an interval of $\pm \$12$.

The use of a formula to obtain an estimate of sample size does not give us more than a rough approximation. In practice it is advisable to take the sample-size estimate as a bare minimum, to be increased for safety.

How can we establish appropriate sample size if we are seeking to estimate a true proportion? What we need, first, is an estimate of P , from which Q follows ($Q = 1 - P$). Here also we may operate on the basis of previous experience or an exploratory small-sample study. If, however, neither of these two approaches is possible, then we can conservatively assume that $P = 50\%$ which leads to a larger sample size than any other value of P .^{*} But suppose that we have reason to expect that $P = 30\%$. Our second need is to decide about precision: we may consider an interval of, say, $\pm 6\%$ around a sample proportion as satisfactory in this situation. Our third need is for a decision about confidence: if we want practical certainty that the true proportion will lie within the interval of $\pm 6\%$, then $\pm 6\% = \pm 3\sigma_p$, and $\sigma_p = 2\%$. If we are satisfied, however, with a 95% degree of confidence, then $\pm 6\%$ equals approximately $\pm 2\sigma_p$ and $\sigma_p = 3\%$.

We are now ready to use the formula

$$n = \frac{PQ}{\sigma_p^2},$$

where $P = 30\%$ or 0.3 and $Q = 0.7$.

At the level of practical certainty, where $\sigma_p = 2\%$ or 0.02,

$$n = \frac{0.3 \times 0.7}{(0.02)^2} = \frac{0.21}{0.0004} = 525.$$

For a 95% degree of confidence we obtain a much smaller sample size, namely,

$$n = \frac{0.21}{0.0009} = 233.$$

^{*} In a 50%-50% break, the numerator PQ in the formula $n = PQ/\sigma_p^2$ is the largest possible and therefore n is largest in such a situation.

The discussion of sample size in this section has referred to unrestricted random sampling. In restricted random sampling, adjustments must be made. In practical work, however, the procedure for approximating sample size as outlined here is often useful as a guide even if restricted random sampling is to be used.

PURPOSIVE SAMPLING

Up to this point our discussions on sampling have referred to random or probability sampling. We now turn, as was promised earlier, to purposive sampling.

The purposive sample is also called the *judgment sample*. These terms indicate selection by design—by choice, not by chance. In purposive sampling a sample is chosen which is thought to be typical of the universe with regard to the characteristic or characteristics under investigation.

For example, in an article on a certain occupational study the *New York Times* reported on June 18, 1950, that the conclusions were reached by "a sampling of 40 representative American colleges and universities." This expression means that 40 colleges and universities were included that appeared to be typical of all American colleges and universities with regard to occupational problems. The sampler has to know certain characteristics of the universe beforehand in order to recognize items in the universe as typical. Another example: if a political analyst believes in the saying "As Maine goes, so goes the Union," he has to know about the political behavior of the country in general; out of this knowledge he declares Maine to be typical of the entire country with regard to political behavior. Thus Maine is a purposive sample; its voting preference is investigated and the findings applied to the universe, namely, the United States. Maine votes in September, the rest of the country in November. (To be sure, Maine has not always proved to be typical of the universe.)

Aspects of sample designs as discussed for probability sampling

can be used in purposive sampling. Thus we find stratified or clustered purposive samples. One type of purposive sample, the *quota sample*, has considerable importance in market and opinion research. A quota sample is basically a stratified purposive sample. The actual selection of sampling units is left to the interviewer, who is told how many units having certain characteristics he is to obtain, but not which ones. The quotas are usually set up in terms of known strata in the universe, for instance, age, sex, occupation. The interviewer is then instructed to obtain certain information from, say, 20 businessmen over 50 years of age. The interviewer is free to select the persons as long as they fall into this group.

Comparison of Purposive and Probability Sampling

The advantages and disadvantages of purposive sampling are best discussed by comparing it in four respects with probability sampling.

1. Since *sampling theory* is based upon chance selection, its principles are not applicable to purposive samples. One consequence of this is that we cannot establish by sampling theory the *reliability* of an estimate of a universe characteristic obtained from a purposive sample. Confidence limits do not apply to purposive samples since the concept of standard error applies only in random sampling. Thus purposive sampling leads to point estimates only and their reliability cannot be gauged by sampling theory.

2. *Sample size* is an important basis for comparing purposive with probability sampling. As we know, large confidence intervals are attached to estimates of universe characteristics made from small probability samples. Where the sample size must be kept small (for reasons such as availability of sampling units or the necessity for keeping costs down), purposive samples are sometimes justifiable.

In probability sampling, an increase in sample size increases the precision of estimates. Such increase in precision does not necessarily follow from an increase in the size of a purposive sample.

The size of the probability sample appropriate for some given situation may be roughly approximated in advance (see pages 441–444). But since the standard error is used in such approximating of sample size, this method cannot be applied to purposive sampling.

3. The *selection* of a purposive sample as distinct from random selection aims at a sample which the sampler deems typical of the universe. This judgment at times requires more information about the universe than is accessible. Furthermore, the human mind has difficulty in recognizing “typical” items. For example, if we were to select “typical” expense vouchers, we would in all likelihood obtain a mean voucher value that differs considerably from the true mean. Thus this difficulty tends to distort purposive sampling.

In quota sampling specifically, the selection of the sampling units is left to the judgment of the interviewer, and here reasons such as accessibility of the sampling units will often determine the make-up of the sample.

4. *Cost and time* are important factors in practical sampling work. Purposive sampling often requires less money outlay and less total time than probability sampling. But cost must be seen also in terms of information purchased per dollar; thus cost must be related to the quality of information obtained.

In practical work, statisticians often use purposive samples successfully in exploratory investigations which precede major survey projects—so-called pilot studies.

OTHER TYPES OF SAMPLING

The Chunk *

There is a distinction between a purposive sample and what is called a “chunk.” A purposive sample is selected by design; a chunk is a fraction of the universe conveniently available for investigation. It is a portion of the universe selected neither by

* See William Edwards Deming, *Some Theory of Sampling*, New York, John Wiley & Sons, Inc., 1950, p. 14.

chance nor by choice but which happens to be at hand. For example, we would call a chunk a readily available batch of expense vouchers separated accidentally from the others, or groups of stores in places where we happen to have resident interviewers.

Sequential Sampling

Since the Second World War a new type of sampling has come to the fore. It is called sequential sampling and it may be characterized as sampling in installments.

Let us illustrate sequential sampling by using the example of expense vouchers, once again. We may be concerned about the computational errors in the 32,000 expense vouchers. We realize that we have to live with some errors of this sort but we must assure ourselves that they are not excessive. We decide that we will accept the entire lot of 32,000 vouchers as sufficiently good for further analysis if 0.4% or fewer are in error but that we will definitely reject the lot if 0.8% or more of the vouchers show computational errors. If errors are found in more than 0.4% and less than 0.8% we will examine further samples. We also agree that we wish to have a 95% degree of confidence in our decision as to the acceptability of the 32,000 expense vouchers. Using these measures we arrive at acceptance and rejection numbers.*

We then begin sampling with an "installment" of, say, 300 vouchers drawn at random. We ascertain the percentage of vouchers that have computational errors in this batch of 300. The percentage of faulty vouchers is compared with the acceptance and rejection numbers for this particular sample size. On the basis of whether the percentage of faulty vouchers in this installment of 300 is (a) below the acceptance number or (b) above the rejection number, or (c) between them, we decide either (a) to accept the entire lot of expense vouchers or (b) to reject the lot or (c) to continue the sampling if no such choice

* See Robert Ferber, *Statistical Techniques in Market Research*, New York, McGraw-Hill Book Company, Inc., 1949, chapter VII.

can yet be made. In this last event, a second installment of 300 vouchers is drawn at random. The percentage of faulty vouchers for both installments combined is established and the above procedure to accept, reject, or continue made on the basis of 600 vouchers. Additional installments are added until a decision can be made to accept or reject.

Three characteristics of sequential sampling become evident from this illustration.

1. Sequential sampling is used in problems involving alternatives, as to accept *or* to reject the entire lot of expense vouchers for further analysis.

2. Unlike all other types of sampling, in sequential sampling the total sample size is not determined at the time sampling starts. We do not know in advance how many installments will be needed until we can arrive at a decision.

3. As was mentioned above, each installment is selected at random. Thus sequential sampling is probability sampling. The major advantage of sequential sampling is that it usually permits the use of a sample of smaller size than would be used in unrestricted random sampling of the ordinary kind. It is not unusual to arrive at a decision after investigating in sequential samples half as many sampling units as would be required in sampling with predetermined sample size.

Sampling units must be capable of being observed in succession; for instance, if observations must be made at one point in time, we cannot sample sequentially. Herein lies a limitation in the applicability of sequential sampling.

Summary

1. Random samples and purposive samples are the two main types of samples.

2. There are unrestricted random samples and restricted random samples.

3. Three important types of restricted random samples are stratified samples, cluster samples, and systematic samples. Each of these three has advantages and limitations.

4. The appropriate size of a sample depends on the individual instance. Sampling theory provides a method of approximating sample size.

5. Purposive samples are not probability samples; they are selected by choice, not by chance.

6. Sequential sampling has special advantages relative to sample size when the problem involves alternatives.

APPENDIX—STATISTICAL QUALITY CONTROL

CONCEPT AND HISTORY

Quality control has been defined as “an effective system for coordinating the quality maintenance and quality improvement efforts of the various groups in an organization so as to enable production at the most economical levels which allow for full customer satisfaction.”*

Quality-control programs involve extensive use of statistical techniques. In the broad sense, therefore, *statistical quality control* means the employment of various statistical principles in quality-control programs. In a narrower sense, statistical quality control is identified by the main statistical tools used—*control charts* and *sampling tables*. Both tools are based on principles of statistical inference (discussed in this book in Chapters 21, 22, 23) and are designed to aid in decisions on the quality of a product.

The word *quality* in statistical quality control refers, for practical purposes, to whatever property of a product must conform to a desired standard.

In statistical quality control, sampling is the basis for establishing the quality of a product. Some form of sampling inspection was

* A. V. Feigenbaum, *Quality Control, Principles, Practice, and Administration*, McGraw-Hill Book Co., New York, 1951, p. 1.

practiced by industry before 1924,[†] but statistical quality control really began in 1924 with Dr. Walter A. Shewhart in the Bell Telephone Laboratories. In the 1930's only a small number of plants, mainly in the electrical and textile industries, had adopted this technique. With World War II, statistical quality control came into its own in the United States; the two major factors in its recognition were its use by the military procurement agencies and its acceptance by scientific societies. By the 1950's, statistical quality control had been accepted, with still further growth assured. One area of such growth is in the application of statistical quality control to certain problems of administration such as clerical and accounting procedures. Thus, statistical quality control has been successfully applied to operations such as key punching and transcription of figures.

A large professional association, the American Society for Quality Control, has since 1946 contributed much to the development of this subject.

Two important areas of statistical quality control are *acceptance sampling* and *process control*. Brief discussions of these will clarify the elements of statistical quality control.

Acceptance Sampling

Acceptance sampling is concerned with the decision to accept a mass of items as conforming to standards of quality or to reject the mass as nonconforming. The decision is reached through sampling. The alternative to sampling is complete enumeration, or what is called 100% inspection, where everything is tested. Then why do we use sampling? Why not inspect every item and make sure? It has been found that 100% inspection does not always mean 100% assurance of conformity to standards of quality.

It is generally true that there are likely to be relatively fewer inspectors' mistakes on sampling inspections than on 100% inspection. As sampling needs fewer inspectors, it often can get better ones. There is a psychological lift in the inspector's knowledge that a decision on an entire lot is to be based on the evidence of the sample he is inspecting; this lift is missing in 100% inspection. And 100% inspection has the serious disadvantage of greater inspection fatigue. Moreover, in many cases 100% inspection is out of the

[†] Sebastian B. Littauer, "The Development of Statistical Quality Control in the United States," *The American Statistician*, October 1950.

question, and the best that can be done is to develop the most satisfactory possible scheme of sampling inspection. Satisfactory sampling acceptance schemes can be developed provided the objective is lowered from the impracticable one of 100% quality assurance to the practicable one of a reasonable assurance that the product accepted is at an acceptable quality level.*

In addition to saving time and money, statistical quality control reduces the number of inspectors on the job and in times of labor shortage this economy in personnel may assume even greater importance than economy in time or money. Moreover, where quality can be tested only by destroying items, as in determining the strength of glass containers, 100% inspection is out of the question and sampling must be used. Of course, there are situations where 100% inspection is not to be put aside; for example, in testing rifles to be used by soldiers, we cannot risk imperfection in any item and therefore must test every one.

In acceptance sampling a so-called lot-by-lot sampling inspection is very often performed. Here a lot is accepted or rejected on the basis of information yielded by sampling. For example, a manufacturer may apply acceptance sampling to purchased raw material and to semifinished and finished products, and through it arrive at a decision to accept or reject a given lot. He makes, as it were, a "yes or no" decision, with no estimate as to the extent of unacceptability of the lot.

Several acceptance sampling plans are in use. Some plans call for one sample of predetermined sample size to be drawn from the inspection lot. Other plans provide for the drawing of two or more successive samples if the first sample does not permit a satisfactory decision as to the acceptance or rejection of the entire lot. If the sampling plan calls for a second sample when the first sample makes for indecision, we speak of double sampling. If sampling is continued until a decision to accept or reject can be reached, we have sequential sampling (see page 448 herein).

Acceptance sampling plans may be distinguished not alone by the type of sampling operation (single, double, and sequential), but also by the types of measurement performed. Where the items forming an inspection lot may be classified as conforming to a given standard (nondefective) or nonconforming (defective), acceptance sampling is

* Eugene L. Grant, *Statistical Quality Control*, McGraw-Hill Book Co., 1946, pp. 356-57; see also p. 9 of the second (1952) edition of Professor Grant's book.

by *attributes*. For example, an industrial product having a wrong color may be considered defective. If in a sample, 5% or more of the items are off color and hence defective, the lot from which the sample was selected may be considered unacceptable.

It may be possible to express a quality characteristic quantitatively; for example, a lot may be considered acceptable if the mean length of the items in the sample is below 2.5 inches but not acceptable if it exceeds this length. Sampling plans of this general type are called acceptance sampling by *variables*.

It may be noted that even where a quality characteristic may be quantitatively differentiated, sampling may be by attributes. For example, if a stated proportion of the items in a sample exceeded a given length, the lot may be considered unacceptable. Thus we may form a twofold classification of defectives and nondefectives, with all items exceeding a given length as the defectives.

In practical inspection work, prepared sampling tables are widely used. These tables give "a series of schedules for representing the probable quality relationships . . . of the entire lot to the samples properly selected from that lot." * Thus in practical work we often make a decision about the lot by reference to prepared tables.

Process Control

Process control, as the name implies, is the application of statistical quality control to the control of a repetitive process of a manufacturing or clerical operation. The units of a manufactured product are never exactly alike; in every manufactured product the quality varies from one unit to another.

Let us illustrate the theory of process control by using compression springs as an example. In the production of an item such as compression springs we find "many small chance variations all the way through the process—in diameter and hardness of the wire itself, in play of the many gears and tools of the machine that makes them, and other similar chance factors. These small chance factors affect each piece separately. Influences in one direction tend to cancel out influences in the other, and the resultant variations in all the pieces of a batch form a definite pattern. . . . For *every* manufacturing process such a pattern of variation *always* exists. But, even more important, for any particular manufacturing process its own

* Feigenbaum, *op. cit.*, p. 188.

pattern will be repeated over and over again, as long as only the same chance causes are at work.”*

Some quality variation due to chance must be expected and is unavoidable. The pattern of variation in quality will remain unchanged within certain limits as long as *only chance causes* of variation are present. But variation in quality may become greater than can be explained by the action of chance. Such variation is an indication of an *assignable cause* of variation. In the production of compression springs, for example, a machine may have gone out of order or an untrained operator may have replaced a trained one.

Statistical quality control provides a way to discover assignable causes of quality variation when they begin to affect a process. If a process goes “out of control,” this fact can be noticed at an early stage. Thus, statistical quality control helps substantially in reducing waste, delays, and the percentage of scrap.

The statistical tool employed in process control is the *control chart*.

CONTROL CHARTS

The Concept

Control charts are simple, routinized devices for checking the pattern of a process. They permit us to keep tabs on a manufacturing or clerical operation while it is in process. Usually small random samples are drawn at different times during manufacture, and inferences are made concerning the universe of products.

We seek to judge whether each given sample is likely to have “come from” the universe of “products in control.” If a sample appears as not coming from the universe of “products in control,” some aspect of the process is out of order and immediate corrective action may be taken.

Thus the control chart may be viewed as a graphic presentation of a running test of statistical significance. The basic statistical concept behind the control chart is the sampling distribution (see Chapter 21 herein).

Let us consider a simple illustration. Before going into the mass production of a special type of nails, a manufacturer produced a

* *An Introduction to Statistical Quality Control*, Bureau of Ordnance, United States Department of the Navy, United States Government Printing Office, p. 7.

number of batches, established the mean length in each batch, and the mean based on all the batch means. This mean of means is, say, 3 inches. Of course he could also have established the mean of one very large batch of nails. The standard deviation for all these nails produced before entering into mass production is, say, 0.22 inch.

These values of 3 inches for the mean and 0.22 inch for the standard deviation may be used as *estimates* of the measures μ and σ (the true mean and the standard deviation respectively).

We are now in a position to construct a sampling distribution of means obtained from random samples of, say, 5 nails (see Illustration A23.1). Its mean is the true mean μ , and the standard error of the mean is equal to $\sigma/\sqrt{n}$. This is, in our illustration, $0.22/\sqrt{5}$ or approximately 0.1 inch.

In statistical quality control we usually consider a sample as coming from the original universe of products "in control" if its mean falls within plus and minus 3 standard errors from the estimated true mean. That is, in terms of our illustration, 3 inches $\pm$ 0.3 inch, or between 2.7 and 3.3 inches. We know that 99.7% of all sample means drawn from the universe of products "in control" will fall within

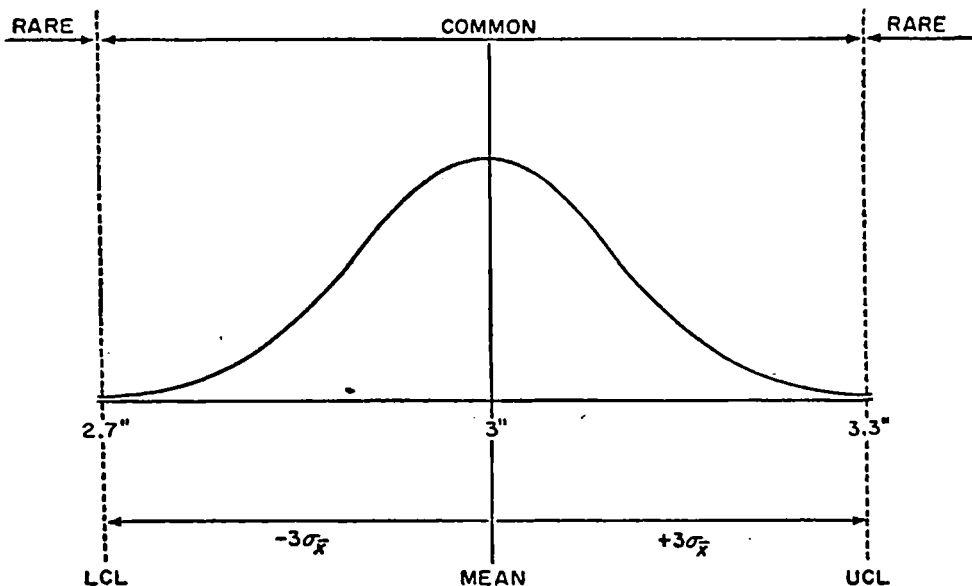


Illustration A23.1

these limits.* Thus, if during the mass production of nails we draw random samples of 5 nails, samples having means smaller than 2.7 inches or larger than 3.3 inches can be presumed to come from a different universe; they indicate that some aspect of the manufacturing process has really changed. An indication of an assignable cause of variation has appeared.

The limits which we set up at a distance of 3 standard errors from the mean of the sampling distribution become control limits and are called *lower control limit* (LCL) and *upper control limit* (UCL).

A control chart can be thought of as a sampling distribution turned around. Illustration A23.1 shows the sampling distribution with μ of 3 inches and σ of .1 inch. Illustration A23.2 (a so-called $\bar{X}$ chart) presents a control chart based on this sampling distribution.†

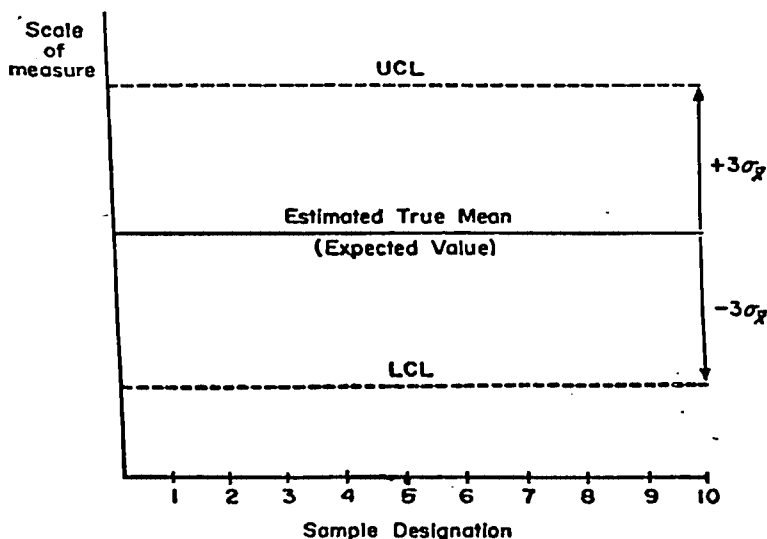


Illustration A23.2

Our illustration refers to the mean of a quality characteristic, namely, the mean length of a special type of nail. In judging the process through the means of the samples taken we obtain an indica-

* This holds true regardless of the sample size since the standard error here is derived from the universe standard deviation as estimated from the standard deviation of a large sample.

† The estimated true mean is often designated by the symbol $\bar{\bar{X}}$ (X double bar).

tion of the level of the process. We usually wish to know, in addition, the uniformity of the process. Accordingly, different types of control charts are used to depict these different aspects.

The concepts of the other types of control charts are analogous to the one described above for the $\bar{X}$ chart.

Types of Control Chart

Control charts may be set up for several statistical measures of a given quality characteristic. The most important statistical measures used for the given characteristic are:

1. Average (measured by the mean $\bar{X}$).
2. Dispersion (measured by the standard deviation s , or by the range R).
3. "Fraction defective" (p) or "number of defectives" (pn).

The arithmetic mean is the average most frequently used. A control chart for $\bar{X}$ is sometimes used alone, but very frequently it is paired with a control chart for either R or s . The range is often used as a measure of dispersion because of its simplicity, but it should not take the place of the standard deviation when samples consist of more than 10 to 15 items.

"Fraction defective" (p) means the fraction of the total number of items in a sample which does not meet the set requirement. "Number of defectives" (pn) is analogous to "fraction defective" but here the *number* of defectives is used instead of a fraction. If only information on "passed" or "failed" is available for a given product (rather than degrees of a quality characteristic), the p or pn control chart is the only possible type.

How Control Charts Are Constructed

In constructing a control chart the following seven considerations are usually involved:

1. The total product is broken up into a number of subgroups. Frequently the basis for establishing subgroups is time; for instance, the production during one hour forms a subgroup.
2. Small samples are drawn from each subgroup; for instance, five items are chosen as a sample of an hour's production.
3. One or more statistical measures are established in each sample

for the quality characteristic under investigation; for example, the mean thickness of 5 items, or their range (that is, the difference between the largest and the smallest of these 5 expressions of thickness), or both.

4. For each statistical measure involved, standard values must be established. These are the expected values which are plotted as the "control lines" of the control chart (see Illustration A23.2). The ex-

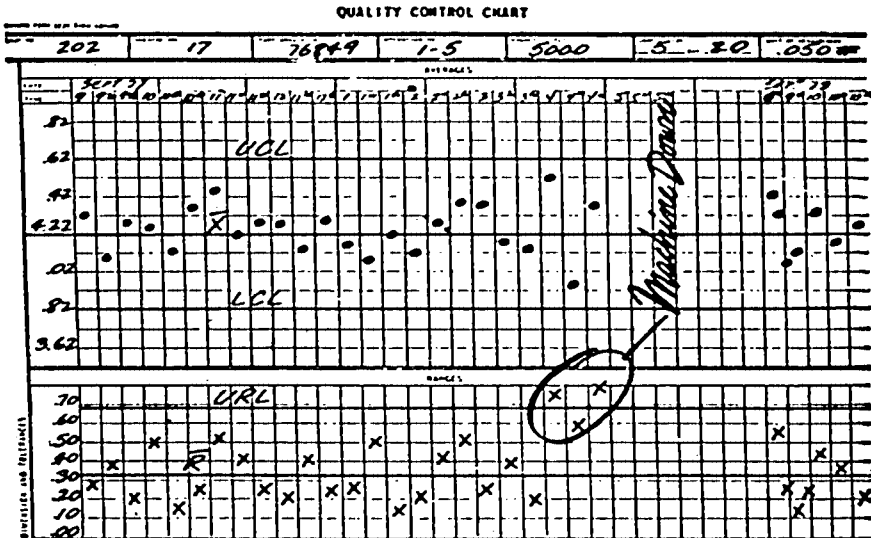


Illustration A23.3

pected value may be, for example, mean thickness, and may be based on past production experience, or on a trial run preliminary to mass production, or on customers' specifications.

5. Control limits must be computed. Tables are available which assist in arriving at control limits.* These control limits are attached to the expected value, thus establishing limits above and below the expected value.

6. A control chart is generally constructed on arithmetic graph

* These tables can be found in books on statistical quality control; for example, Eugene L. Grant, *Statistical Quality Control*, McGraw-Hill Book Co., Inc., 1952, and Acheson J. Duncan, *Quality Control and Industrial Statistics*, Richard D. Irwin, Inc., Homewood, Ill., 1953.

paper. Separate presentation is made for each statistical measure chosen; for instance, for $\bar{X}$ and for R (see Illustration A23.3).

The horizontal scale shows the numbers assigned to the successive samples, but time (such as the hour of the day) may be indicated instead. Along the vertical scale are the values for plotting the observations.

Control limits are customarily plotted as broken lines; the "line of expected value" or "central line" is a solid line. If control charts for two different statistical measures of the same characteristic are shown, then the two horizontal axes must correspond.

7. The values established for the successive samples are now plotted on the control chart. The observed values are represented by dots or circles or crosses, which may or may not be connected by a line.

How to Read a Control Chart

If a control chart is properly constructed, we may draw conclusions from the positions of the dots representing the observed values. Generally, if all the dots remain between the control limits, then we may say that the process is "in control." If dots are outside of the control limits, the process is "out of control." The upper and lower control limits are those limits within which variations generally may be attributed to chance. If all dots are found between the upper and lower control limits, we may for all practical purposes *assume* that only chance causes are present. Dots are, however, sometimes found arranged in some peculiar way. Although they appear between the control limits, a substantial number of successive dots may be located on the same side of the central line, the line of expected value. Or successive dots may follow a definite path leading toward the upper or lower control limit. Such patterns of dots within control limits are danger signals which may indicate a change in the production process.

If one or more dots fall outside the upper or lower control limits, we assume the presence of "assignable causes" and consider the process "out of control." When an observed value is found outside of the control limits, action can be taken promptly to bring the process under control.

An Illustration of the Use of a Control Chart

Let us illustrate the application and interpretation of an $\bar{X}$ and R Control Chart as used by the Bureau of Ordnance of the United States Department of the Navy for testing the production process in the manufacture of compression springs.

"Samples were taken as rapidly as possible from the early part of the run and results of the test were plotted on the control chart.

"By analyzing the control chart the correct machine settings were rapidly found and assurance was given that process variability could be kept within the limits of the Navy specifications. This was different from the old days in the plant when the machine operator, after initial setup, was left to do the best he could, and how well he succeeded was not known until final screening inspection.

"The process remained in control for several days, but one day a sharp increase was noticed in the sample ranges plotted on the control chart. It was quickly estimated that the machine would now produce about 20% scrap. In spite of thorough checks of the tools and their settings, the operator and the department head could not spot any visible cause of the difficulty.

"Technical investigation of current output of the machine quickly revealed that the variation was cyclical, occurring once about every three springs.

"The cyclical nature of the variation seemed to rule out as the source of the trouble those parts of the machine which operate once for each spring produced. But the investigation did suggest that the trouble might be found in the feed rolls, which for this particular spring made about one revolution for every three springs produced. Even though the feed rolls seemed satisfactory visually, they were dismantled and carefully checked. It was then discovered that the shaft of one was slightly out of line. With the shaft trued up, the operation was immediately brought back into control and production continued smoothly till the order was filled." *

Illustration A23.3 depicts the situation described above. $\bar{R}$ means average range, which is the expected value of the range chart, and URL or upper range limit is sometimes used instead of UCL in this type of chart. The lower limit in this illustration of a range chart is zero.

* *An Introduction to Statistical Quality Control*, Bureau of Ordnance, United States Department of the Navy, United States Government Printing Office, p. 16-17.

APPENDIX—RANDOM NUMBERS AND THEIR USE

A random sample may be obtained by using a Table of Random Numbers as below. First, all items in the universe are numbered. The items are taken consecutively regardless of how they are arranged. Let us assume the universe consists of 8000 employees, arranged alphabetically. We number these employees consecutively from 1 to 8000. If we wish to draw a sample of 500 employees, we make use of the Table of Random Numbers (page 462) in the following manner:

Since the universe consists of 8000 items, we use only four-digit numbers. We may decide to start at the top of Column (1) and proceed vertically down the column. We might also start at some other point, or we might proceed horizontally. The Table of Random Numbers may be used in any systematic way.

We read 500 four-digit numbers from the table, and include in our sample the employees designated by these numbers.

In reading these four-digit numbers from the table we disregard any number higher than 8000, since we have only 8000 employees. We disregard all repetitions of numbers.

Thus, from the table on page 462 we would include in our sample the employees bearing the following numbers: 1328, 2122, 19, 6057, 3524, 3898, 1075, and so on.

Line	Col.	(1)	(2)	(3)	(4)	(5)
101		13284	16834	74151	92027	24670
102		21224	00370	30420	03883	94648
103		99052	47887	81085	64933	66279
104		00199	50993	98603	38452	87890
105		60578	06483	28733	37867	07936
106		91240	18312	17441	01929	18163
107		97458	14229	12063	59611	32249
108		35249	38646	34475	72417	60514
109		38980	46600	11759	11900	46743
110		10750	52745	38749	87365	58959
111		36247	27850	73958	20673	37800
112		70994	66986	99744	72438	01174
113		99638	94702	11463	18148	81386
114		72055	15774	43857	99805	10419
115		24038	65541	85788	55835	38835
116		74976	14631	35908	28221	39470
117		35553	71628	70189	26436	63407
118		35676	12797	51434	82976	42010
119		74815	67523	72985	23183	02446
120		45246	88048	65173	50989	91060
121		76509	47069	86378	41797	11910
122		19689	90332	04315	21358	97248
123		42751	35318	97513	61537	54955
124		11946	22681	45045	13964	57517
125		96518	48688	20996	11090	48396
126		35726	58643	76869	84622	39098
127		39737	42750	48968	70536	84864
128		97025	66492	56177	04049	80312
129		62814	08075	09788	56350	76787
130		25578	22950	15227	83291	41737
131		68763	69576	88991	49662	46704
132		17900	00813	64361	60725	88974
133		71944	60227	63551	71109	05624
134		54684	93691	85132	64399	29182
135		25946	27623	11258	65204	52832
136		01353	39318	44961	44972	91766
137		99083	88191	27662	99113	57174
138		52021	45406	37945	75234	24327
139		78755	47744	43776	83098	03225
140		25282	69106	59180	16257	22810
141		11959	94202	02743	86847	79725
142		11644	13792	98190	01424	30078
143		06307	97912	68110	59812	95448
144		76285	75714	89585	99296	52640
145		55322	07598	39600	60866	63007
146		78017	90928	90220	92503	83375
147		44768	43342	20696	26331	43140
148		25100	19336	14605	86603	51680
149		83612	46623	62876	85197	07824
150		41347	81666	82961	60413	71020

Random Numbers, from *Table of 105,000 Random Decimal Digits*, Statement No. 4914, Interstate Commerce Commission, Bureau of Transport Economics and Statistics, page 3.

PART
5

Misuse

Misuses of Statistics

The Problem of Misuse

Like any other part of scientific method, statistics is capable of misuse. Since statistics bears on the daily concerns of the citizen, the businessman, the economist, the government, and its use is so widespread, it is vitally necessary to be able to recognize fallacies in its application. Not only must we recognize the misuses as consumers of statistics, but in addition, a study of them will help us avoid them as producers of statistics. Now that sound statistical method has been presented, pitfalls can be more easily recognized.

The motive or intent of the misuser of statistics is not here at issue; fallacies designed to mislead and those committed unintentionally will not be distinguished. The effect is the same in both cases—although, to be sure, in the first case there is an abuse of statistics, in the second case only a misuse.

We are not concerned with sheer computational and mechanical errors, but with errors in collection, in presentation, in analysis, and in interpretation. Some of the illustrations in this chapter bear on more than one of these four parts of statistical work, but we have classified each one according to what we consider the major element of misuse.

Some of the material in this chapter we have adapted from other authors.*

MISUSES IN THE COLLECTION OF DATA

Incomparable Data

A recruiting officer for the United States Navy claims that it is safer to be a sailor in the Navy during a war than a civilian in a large city such as New York. He bases his claim on the fact that during the Spanish-American War the death rate in the Navy was 9 per 1000 while for the same period in New York City the death rate was 16 per 1000. But the two groups are not identical as to age, sex, and health, for the New York City death rate includes the mortality of infants, old people, and people in hospitals and asylums, while the Navy is mainly composed of men between the ages of 18 and 35 who are in the pink of physical condition.

A sound comparison would require collecting data on the death rate of noncombatants between the ages of 18 and 35, in the pink of condition, who are residents of New York City; or at the very least of all noncombatants in New York City between the ages of 18 and 35 regardless of physical condition.

Failure to Consider Changes in Classification

It was claimed that there was a back-to-the-farm movement from 1930 to 1935 because the United States Census showed 500,000 more farms in 1935 than in 1930. But the method of

* Robert E. Chaddock, *Principles and Methods of Statistics*, Houghton Mifflin Company, 1925. Jerome B. Cohen, "The Misuse of Statistics," *Journal of the American Statistical Association*, vol. 33, No. 204, December 1938, pp. 657-674. Morris R. Cohen and Ernest Nagel, *An Introduction to Logic and Scientific Method*, Harcourt, Brace and Company, 1937. Harold A. Larrabee, *Reliable Knowledge*, Houghton Mifflin Company, 1945. Horace C. Levinson, *The Science of Chance*, Rinehart and Company, 1950. Elmer B. Mode, *The Elements of Statistics*, Prentice-Hall, second edition, 1951. William A. Neiswanger, *Elementary Statistical Methods*, The Macmillan Company, revised edition, 1956. Raleigh Schorling, John R. Clark, Francis G. Lankford, Jr., *Mathematics for the Consumer*, World Book Company, 1947.

applying the Census definition of a farm had been changed between 1930 and 1935. It was estimated that the 1930 local enumerators in farm areas had failed to report 300,000 farms which would have been included had the definition been literally interpreted that year. Consequently, it is hardly possible to say whether the increase of 500,000 farms in 1935 over 1930 indicates back-to-the-farm movement or is a result of the change in collection of the data.

Biased Sample

The two most famous instances of this misuse in the collection of data relate to two presidential campaigns in recent United States history, that of 1936 and that of 1948. In 1936 the *Literary Digest*, a weekly magazine no longer published, predicted the election of the Republican Landon over the Democrat Roosevelt. Yet on Election Day Roosevelt carried every state in the union except Maine and Vermont. What was wrong with the *Literary Digest* poll? Straw-vote ballots had been sent to ten million people by the magazine. Of those who sent in their ballots in this 1936 poll, 1,078,012 had voted the Republican ticket in 1932 and 1,020,010 had voted the Democratic ticket in 1932. Since in 1932 there had been 22,822,000 Democratic voters as against 15,762,000 Republican voters, the magazine's sample was "biased" in favor of the Republican candidate. That is, the percentage of Republican answers to the poll was far higher than the percentage of Republicans in the population at large, based on the 1932 figures. In a fair sample of 1936 voters, the same proportion between Republicans and Democrats should have been maintained as appeared in 1932. Furthermore, the questionnaires were sent only to owners of automobiles and telephone subscribers, and not every voter is an owner of an automobile or a telephone subscriber. Thus, only a certain socio-economic stratum was polled.

In November 1948, all the public-opinion polls predicted that Harry S. Truman would lose and Thomas E. Dewey would win. One poll even stopped interviewing voters in September, two

months before election. But a swing to Truman occurred in the meantime, either from the "undecided" voters in the poll or from "Dewey" voters. In addition, there was no assurance that those polled would really vote on Election Day; and there is some evidence that Republican voters stayed away, assured of victory by the very polls which were disproved. "Many voters simply do not know how they will vote until they arrive at the voting booth. When asked how they intend to vote, they indicate one of the candidates or say they are undecided, depending on their temperaments and habits of thought. . . . It is safe to say that no polling technique based on small samples will ever predict with accuracy a close election, nor will it predict with accuracy a fairly close election in which there is a rapid shift in sentiment just before election day."*

Incomplete Enumeration

In China, a regional census taken for poll-tax and military purposes returned a population of 28,000,000. A few years later, when population statistics were gathered over the same territory, this time however for famine relief, the population "increased" to 105,000,000. The figures for the first of these censuses were obviously inaccurate, owing to the reluctance of people to expose themselves either to the poll-tax or to military duty.

MISUSES IN THE PRESENTATION OF DATA

Failure to Present Complete Classification System

It was reported that during the year 1946 the number of deaths chargeable to steam railroads was 4712. This is a very high figure and one likely to make people think that steam railroads are not safe means of transportation. But complete classification of these deaths results in a far different picture. Of the 4712 only 132 were passengers on railroad trains; 736 were railroad em-

* Levinson, *op. cit.*, p. 268.

ployees; 1618 were trespassers on railroad property (mostly those "riding the rods"). And deaths at grade crossings were 2025. There were 201 deaths classified as "other nontrespassers."

Prior to our entry into World War I, the *Navy Year-Book*, in keeping with the desires of the advocates of preparedness, counted obsolete and obsolescent vessels in listing the navies of Germany and other countries but omitted them in the case of the United States, and also counted certain German and other battle-ships, building or provided for, but omitted similar ones for the United States and thus made our Navy appear inferior.

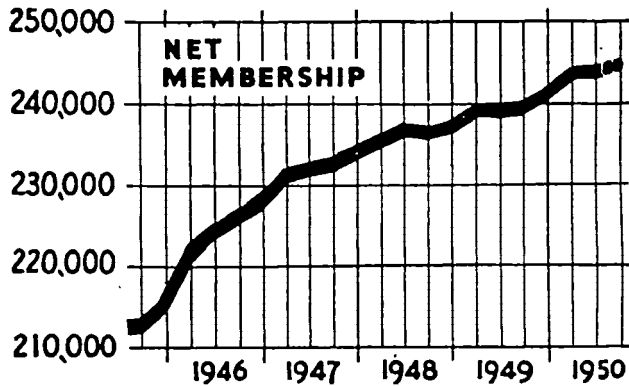


Illustration 24.1. No Zero Line.

Spuriously Accurate Presentation

On February 9, 1949, it was reported in the newspapers that a New York State study showed that a working woman living with her family needed a weekly salary of \$40.13 for adequate support. This seems to show that we are able to gauge to the nearest penny how much such a woman needs to keep body and soul together. The very large mass of statistics involved in arriving at such a figure precludes such accuracy. It is far preferable to say that a working woman living with her family needs about \$40 for adequate support.

The American Petroleum Industries Committee reported that

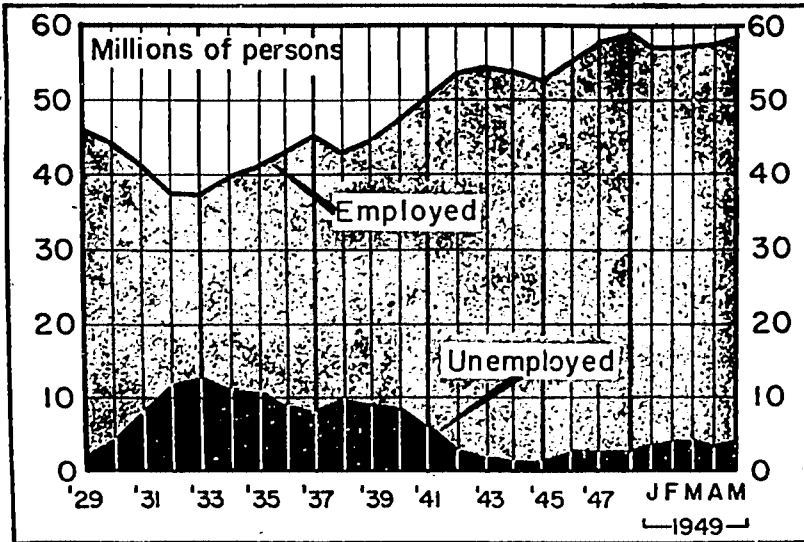


Illustration 24.2. Failure to Represent Equal Magnitudes by Equal Spaces on the X-Axis.

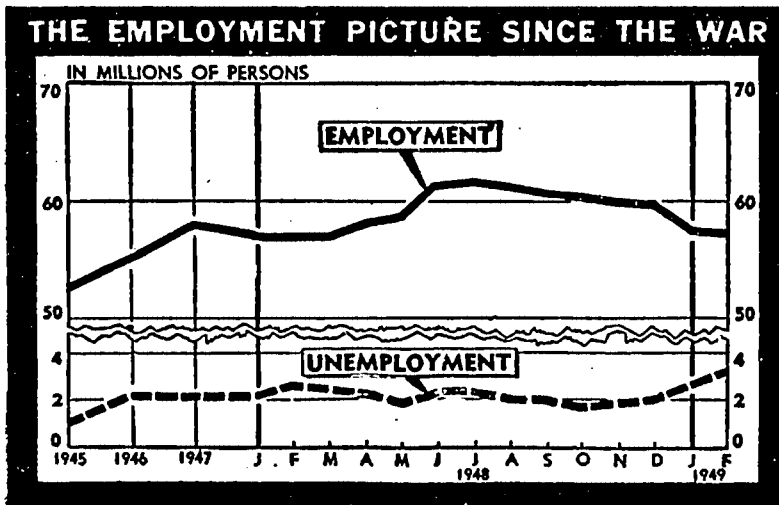


Illustration 24.3. The Use of Two Different Vertical Scales on Arithmetic Grid Is Likely to Give False Visual Impressions.

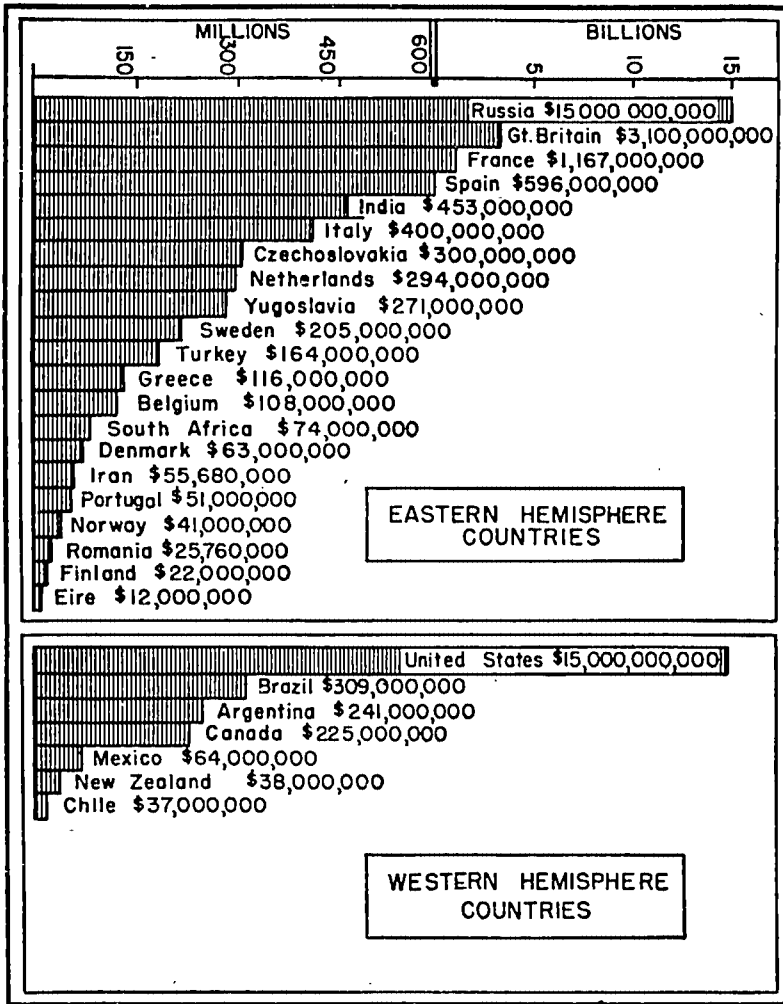


Illustration 24.4. Two Different Scales on the Same Axis.

the average tax bill per motor vehicle in the United States for 1938 was \$51.13. There is great difficulty in finding out exactly how much the individual motorist pays in various taxes, and estimates are used in some instances. Hence, the result has been stated with unwarranted precision.

Errors in Graphic Presentation

Illustrations 24.1–24.6 of faulty uses of graphic presentation have been taken from newspapers, periodicals, and advertising. We have not quoted the sources here.

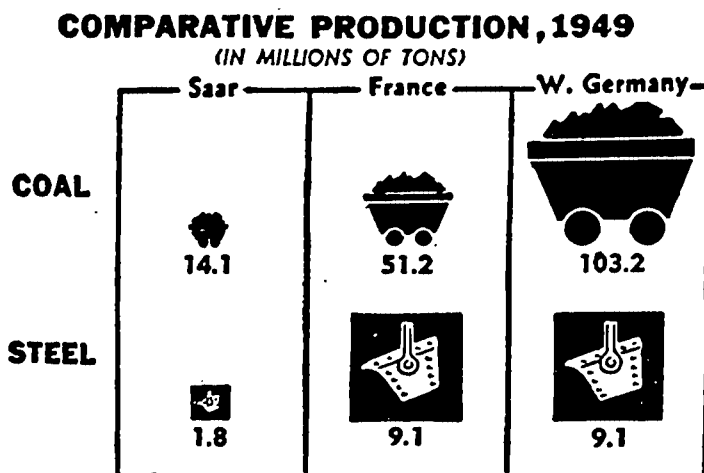


Illustration 24.5. Larger Quantities Are Better Shown by Using a Larger Number of Symbols than by Using Larger Single Symbols.

MISUSES IN THE ANALYSIS OF DATA

Use of Absolute Numbers Instead of Percentages

The author of a book on socialism tried to show that small business, instead of declining as had been predicted by certain socialists, was actually increasing. He used the following data relating to business in Germany in the late nineteenth century.

<i>Commercial Establishments</i>	<i>Number in 1882</i>	<i>Number in 1895</i>
With no employees	429,825	454,540
With 1 to 5 employees	246,413	450,913
With 6 to 50-employees	26,531	49,271
With over 50 employees	463	960

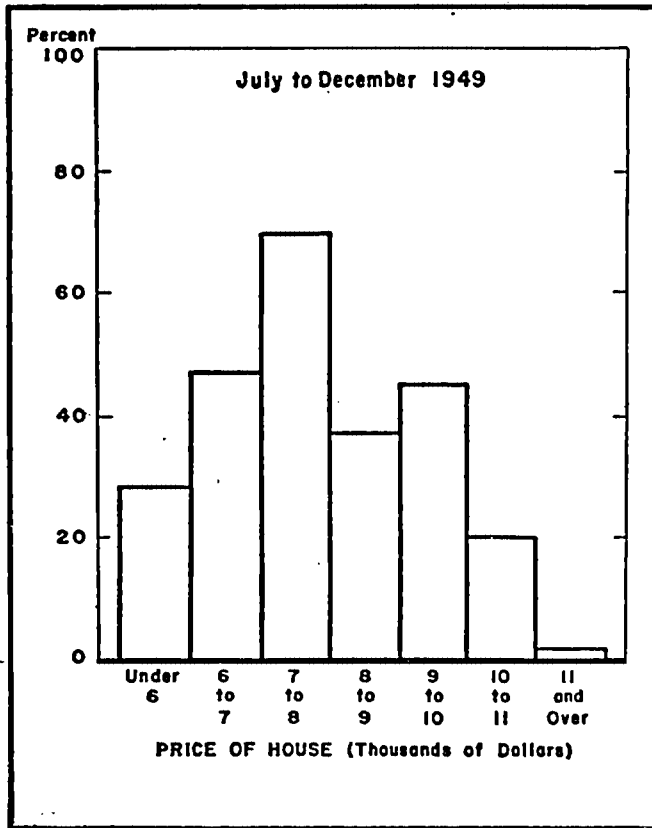


Illustration 24.6. Histogram Cannot Be Constructed in Open-End Distributions, and Class Limits Should Not Overlap.

But if percentages are computed for each of these two periods, the following appears, which does not prove the author's point at all.

	Percent in 1882	Percent in 1895
<i>Commercial Establishments</i>		
With no employees.....	61.12	47.56
With 1 to 5 employees.....	35.03	47.18
With 6 to 50 employees.....	3.78	5.16
With over 50 employees.....	.07	.10

Use of Percentages Instead of Absolute Numbers

Johns Hopkins University some decades ago permitted women to enter classes, thus breaking a precedent. Shortly thereafter, it was reported that $33\frac{1}{3}\%$ of the women students at Johns Hopkins had married members of the faculty. This looked like a sizable number until it was discovered that there were only three women students enrolled there at the time and one had married a faculty member.

Faulty Use of Percentages

The author of a letter to the *New York Times* in 1947, commenting on an article titled "Man's World—or Woman's," said that he trusted "that statisticians will find that women give birth to 100% more babies than men." Since men do not give birth to any babies, we would therefore find that neither do women, since 100% of zero is still zero.

It was reported that in 1939 the value of the Bolivian peso had dropped more than 500%. No number can drop more than 100% and remain a positive number.

Misuse of the Mean

A corporation was controlled by three individuals who owned 1,500,000 shares of the total of 2,000,000 shares of stock; it had 3000 other stockholders who each owned a small number of shares. The corporation announced that the average holding of stock of their stockholders was 666 shares or $2,000,000 \div 3003$.

Failure to Use the Weighted Mean

In a factory, there were 4 heavy-duty maintenance workers who averaged 210 pounds in weight. There were 10 light-duty machine operators who averaged 150 pounds in weight. It was erroneously reported that the average weight of all the fourteen workers was 180 pounds or $(210 + 150) \div 2$. But the weighted mean is

$$\frac{(210 \times 4) + (150 \times 10)}{14} = 167.1 \text{ pounds.}$$

Faulty Use of the Median

If, on an aptitude test given to the workers in a plant, 50% of the workers received a grade of 10% or less, the statement that the median is 10% tells us nothing about the scores of the remaining 50%. Moreover, if in another group of workers taking the same aptitude test, half of them received a grade of about 25% and half a grade of about 85%, the median would be about 55% and would not be informative as to the distribution of the scores on this test.

Faulty Use of the Mode

A list of best-selling books is sometimes made up by investigating a number of stores reporting. The best seller is then stated as the book which most stores report selling most of, which is a modal value. But no account is taken of how many copies of each book reported have been sold by each store. The mode is not appropriate, but the weighted mean should be used to bring out the relative number of sales at each store.

Faulty Use of the Range

The range of wages in a plant is from \$40 a week to \$100 a week; the president of the corporation earns \$500 a week. A report that earnings of employees of the corporation range from \$40 to \$500 a week is deceptive.

Failure to Use a Measure of Dispersion

During a military campaign, the supply corps of an army in the field may report the average food consumption of the soldiers. But if many of its men are short of food through being cut off from supply lines while others are well supplied, the average alone is misleading. Consequently a measure of dispersion must accompany the average; otherwise the average looks highly representative of per capita food consumption without showing the wide scatter.

Faulty Extrapolation of Trend

Abraham Lincoln in his second message to Congress forecast what the population of the United States would be in 1930. He said that it would be 251,689,914. This prediction he based on the Census figures for 1790 to 1860, which showed an increase from 3.9 million in 1790 to 31.4 million in 1860. By 1930 the population of the United States turned out to be 122.8 million. Lincoln's estimate rested on the assumption that the population of the United States would continue to increase at the same rate as it had from 1790 to 1860.

Extrapolation of trend can be done with some confidence only if other conditions can be shown to remain the same, and over long periods they are not likely to do so.

Faulty Use of Indexes

By using depression lows as a base period, one of the national organizations of labor tried to show that the index of production and the index of profits rose far more rapidly during economic recovery than did employment and wages. On examination of the facts, it turned out that production and profits had fallen much more than employment and wages. Production and profits could rise, percentagewise, more than employment and wages since the base from which the former rose was much lower than the base from which the latter rose.

Misuse of Correlation

It has already been shown that a high coefficient of correlation can be found when no substantive relationship exists between the two variables. It is also possible to find a coefficient of correlation based upon a faulty assignment of dependence to one of the variables. One might observe a correlation between high-speed typing and shortness of the typists' fingernails, but either variable might be dependent—the speedy typing might tend to keep the fingernails short, or the short fingernails might facilitate speedy typing. Or, it may be that there is not a relation of

dependence of one upon the other, but that they mutually react.

Moreover, a linear correlation that is found may exist only within limits. If an agricultural crop is larger the more rainfall there is, that does not mean that the greater the rainfall the larger the crop, for too much rainfall may ruin the crop. Thus, instead of positive linear correlation, we have curvilinear correlation with a positive relationship up to a certain point, and then a negative relationship setting in after that point.

MISUSES IN THE INTERPRETATION OF DATA

Failure to Comprehend the Total Background of the Data

From a study of the mortality statistics on cancer, the interpretation has been made in the press and in popular lectures that the death rate from cancer is increasing. It is true that the number of deaths recorded from cancer has increased greatly in the last quarter of a century. In order to interpret the figures correctly it is necessary to comprehend the total background of the data in terms of the following considerations: (1) Some deaths formerly listed as "causes unknown" are now correctly diagnosed as cancer deaths; (2) due to increasing autopsy analysis of cause of death, there has been an increase in postmortem diagnoses of cancer which show that the cause of death originally reported is not the true cause; (3) there has been great improvement in the reporting of mortality statistics and in the collection of such statistics; (4) cancer is largely a disease of later life and our population has been aging, consequently there has been an increased proportion of the population susceptible to cancer; (5) other diseases heretofore causing death have been brought under control, and consequently many more die of cancer now who would have died of other diseases in earlier times.

In short, more than the bare enumeration of cancer deaths is required to make correct interpretation here.

Interpretation Based on Individual Cases Instead of Average

Since wages in cigar and glass manufacturing were about the same in the South and in the North, it was concluded that in general there was no wage differential between Northern and Southern workers. But if all industries are considered, it is found that earnings of Southern workers average considerably less than those of Northern workers. Here interpretation must be based upon the over-all average.

Interpretation Based on Average Instead of Individual Cases

What is true on the average may not be true for each of the items which make up the average. For example, a denial that *all* American railroads were overcapitalized was supported by showing the mean capitalization of American railroads. But the mean capitalization lumped together solvent and insolvent railroads, and failed to show that some railroads were on the verge of bankruptcy at various times because of overcapitalization.

Confusion of Averages

Interpreting the Mean as if It Were the Median. The mean of human intelligence in the population has been taken; accordingly, "average" intelligence has been affected by the presence of extremes. The low extremes are very prevalent, and pull the "average" down. It has been said there are as many persons above the average as below the average. This statement can be made only on the basis of the median, not the mean.

Mistaking the Mode for the Mean. This common error results from assuming that the most frequent value in a series is the same as the computed average, namely, the mean. Thus the mean grade was claimed to be 80% on a certain examination, whereas 80% was really the most frequent grade or the mode, and the mean was 72% in the case of these grades.

Interpreting Seasonal Variation as Cyclical Variation

Retail sales were higher in April 1938 than in April 1937, and this was cited to prove that economic conditions were improving. But in 1937 the Easter holiday fell at the end of March and in 1938 in the middle of April. Thus, seasonality, not the cyclical factor, caused the April increase in 1938 as against 1937.

Interpreting Cyclical Variation as Seasonal Variation

It was claimed that the increase in Christmas business in 1939 as against 1938 was the result of making Thanksgiving Day one week earlier, thus prolonging the Christmas shopping season. This statement might have been true in part, but it neglects the fact that 1938 was a "recession" year while 1939 was a "recovery" year, and therefore the cyclical factor was also responsible for the better business in 1939.

Interpreting Seasonal Variation as Trend

At the beginning of December 1948, meat prices fell and it was interpreted by many as the beginning of a downward trend. But the commissioner of markets in a large city correctly pointed out that meat prices usually fall in December because of a falling off in demand for meat at this time of year, and that they rise again shortly thereafter.

Time Sequence Interpreted as Causation

Some years ago the United States concluded a reciprocal trade agreement with Canada. Not long after the conclusion of the agreement, imports of foodstuffs from Canada rose sharply. This rise was interpreted as being caused by the letting down of tariff barriers, and as contrary to the interests of United States farmers. But what actually happened was that the droughts of 1936 cut down United States domestic production so sharply that heavy imports were needed. Thus imports followed the trade agreement in time but were not caused by it.

Misinterpretation of the Coefficient of Correlation

The coefficient of correlation can be used to measure the covariation of two series which are really independent of each other. It has been shown that budgetary expenditures for the British Navy are correlated with the consumption of bananas, and the spread of a certain disease in England with the increase in imports of apples. Covariation alone cannot be interpreted as manifestation of functional relationship or causation.

CONCLUSIONS ON MISUSES

The misuses of statistics illustrated here are, unfortunately; far from exhaustive.* What we have presented are certain of the basic types of misuse. One who has learned the basic principles of statistics, and has been shown the "crimes" committed in its name, should now be able to avoid waywardness as a producer of statistics and also recognize it when exposed to it as a consumer of statistics.

We may appropriately close this chapter with the statement of an eminent mathematician. "The abuses of statistics should not blind us to their effectiveness in studies of population changes, stock-market operations, unemployment, wage scales, cost of living, birth and death rates, extent of drunkenness and crime, the distribution of physical characteristics and intelligence, and the incidence of diseases. Statistics are the basis of life insurance, social-security systems, medical treatments, governmental policies, and the numbers racket. Even the hard-headed businessman uses statistical methods to locate his best markets, to control manufacturing processes, to test the effectiveness of his advertising, and to gauge the interest in a new product. The statistical approach eliminates haphazard guesses and the captiousness of individual judgments and replaces them by highly useful conclusions."†

* For a humorous treatment of statistical misuses, see Darrell Huff and Irving Geis, *How to Lie with Statistics*, W. W. Norton & Company, Inc., 1954.

† Morris Kline, *Mathematics in Western Culture*, New York, Oxford University Press, 1953, p. 357-8.

General Appendixes

Appendix

I

How to Make a Statistical Report

Statistical investigations and research seek to solve problems through the collection, presentation, analysis, and interpretation of relevant quantitative data. But the findings of research must be communicated to others. Description of the processes of communication of findings may be separated, as below, into two parts: (1) preparation for making the report; (2) presentation of the report. In our preparation of a report, its ultimate presentation must be the guide. Thus, many problems appear in preparation that also appear in presentation, but on different levels.

Preparation of the Report

In preparing a statistical report, it is essential that the aim and purpose of the work be clearly defined and followed. Records of how the data are collected must be kept so that methods of collection can be stated in the report. Moreover, if sources are used, clear records of them will facilitate the writing of the report. The tabular and graphic presentation must emphasize the significant elements of the research,

and the administrator who is in charge of presentation of tables and graphs must plan precisely what is to be conveyed to the reader. Much good statistical work has been spoiled by poor planning of tabular and graphic presentation.

Throughout the analysis of the data, records should be kept on why certain measures were used. The significance of the classification system, of the averages and measures of dispersion, or of the methods of time-series analysis, will be carefully explained in the report, and therefore the reasons for using the particular techniques should be noted while the report is being prepared.

Preparation must also look toward communication of the interpretation of the data. While statement of the analytic techniques employed communicates to the reader how the data were taken apart, from the interpretation of the data in the report the reader should be able to understand the meaning of the results obtained through analysis. The interpretation of the data leads to conclusions and findings. These should be prepared for statement in a general way. Wherever possible, a summary of findings and conclusions should be prepared. A list of recommendations may be prepared if the scope and program of the research call for it.

Presentation of the Report

The level of statistical understanding of those to whom the report is addressed determines the language to be used and the amount of detail to be included.

Every report should have a title page which also includes the name of any sponsoring agency as well as the organization or individual that prepared it. This title page should be followed by a full and complete table of contents. Any acknowledgments should appear either in a short separate statement or as part of a foreword. The foreword also may give the reasons for making the study.

Frequently, the body of the report opens with a short summarization of results of the entire study.

The body of the report should then state in complete detail the purposes of the investigation. This statement is followed by a full explanation of the statistical methods employed in collecting, presenting, analyzing, and interpreting the data. If sampling was done, details on sampling procedures should be included. The main part

of this section consists of discussion based on analysis accompanied by pertinent tables and graphs.

The statistical findings are then fully described and explained. Finally, the body of the report contains any recommendations.

As a rule, the report should be arranged in such a way that general findings are stated first. Then the report should progressively proceed to the more specific information, ending with the most specific.

In the last part of the report are placed the tables and graphs which were not used in the body of the report. They may be important as reference material on the problem. A copy of the registration form or schedule or questionnaire used to record the information should be included. If collection was by personal interview, instructions to interviewers may be reproduced here.

The major faults that may appear in the writing of the statistical report are inaccurate statement, superfluous and unnecessary language including "overwriting," technical level of communication beyond the capacity of the readers, and poor grammar.

The technical aspects of presenting tables and graphs have been treated in the text of this book in Chapters 4, 5, and 6. Use of such technical aids as lettering devices, printed symbols for pictographs, and the like, improves the appearance of the report and makes the work easier.

In typewritten reports, different levels of heading (chapter titles, subtitles, paragraph titles) should be used and kept consistent throughout. The varying emphasis may be achieved by centering, capitalizing, underlining, or combinations of these.

On large research projects in statistics, the making of the report is given over to a special report writer who has shown particular aptitude for this type of work. This specialization indicates the great importance which must be attached to communicating the results of statistical labors.

Both the preparation and the presentation of a statistical report manifest the qualities possessed by the statistician: technical skill in handling statistical methods; breadth and depth of knowledge of the subject matter; thoroughness and patience; accuracy; imagination and the ability to adapt general statistical principles to the specific situation dealt with in the report; and, finally, impartiality and willingness to communicate the truth regardless of personal feeling and belief.

Appendix

II

Approximate Numbers and Rounding

The statistician must often deal with approximate data, and must use his good sense and perhaps certain rules of scientific computation in dealing with them.

Ratios, for example, are often approximate. It is quite usual to get some such ratio as $97/91$, which might be the ratio of hourly wages paid to certain workers in two factories. This ratio equals $1.06593 \dots$ and all the further computation in eternity will not make it "come out even." So the ratio is approximate, no matter where computation ceases.

It might seem that computing this ratio to five decimal places was careful and precise work, but this refinement is almost entirely useless and might be misleading as a false pretense of accuracy. If this ratio were applied to a salary adjustment affecting the two groups of workers, the figure 3 in the fifth decimal place would make less than a penny's difference in a hundred-dollar pay envelope. So the statistician takes his practical problem into account and computes his ratio only to useful degrees of refinement. The ratio $97/91$ might be used as a

common fraction, or converted to a decimal as 1.07, or 1.066 if the statistical problem required meticulous care.

Another example of approximate data is frequently encountered in discussions of the national income, as when a committee of Congress is planning new income-tax rates. We see money amounts stated in such form as "Total factory wages for next year are expected to increase by \$11 billion over total factory wages for the current year." It would be nonsense in such a prediction to suppose that "\$11 billion" means "\$11,000,000,000, not \$11,000,000,100 and not \$10,999,999,900." The difference of \$100 in \$11 billion would not be equivalent to 1 cent in a million dollars. The approximation "\$11 billion" is good enough for the study of the problem being contemplated.

We have, then, both small and large approximate numerical data. We had the three approximate ratios 1.07, 1.066, and 1.06593; and we had the approximate income rise of \$11 billion. Of these four approximations, we say that 1.06593 was computed to the fifth decimal place; we say of 1.066 that it is the same ratio *rounded* to the nearest thousandth; we say of 1.07 that it is the same ratio rounded to the nearest hundredth; and we say of the \$11 billion that it is rounded to the nearest \$1 billion.

The statistician, in working with approximate data, must decide where to round them. In data of business and economics, common sense will usually tell him where to round a number. In some refined work in the natural sciences the criteria for rounding are more rigorous.

Rounding, however, presents certain practical problems to the statistician in economics and business. He frequently gets data from his sources in the form of round numbers, often of numbers rounded to different degrees of refinement. He may learn from one source that families in Mill City spend 5.4% of their income on medical care and from another source that families in Factoryville spend 6% of their income on medical care. Then he must ask whether the one source rounded its figures to the nearest .1% and the other rounded its figures to the nearest 1%. Unless both used the same practice in rounding, the two percentages cannot be validly compared. Before comparing them, the statistician must make sure that they were rounded alike. The Factoryville figure of 6% might have been rounded from as little as 5.6% or as much as 6.4%; and the ratios at these extremes would be very different. The ratio $5.4/5.6$ is equal to .96 approximately; the ratio $5.4/6.4$ is equal to .82 approximately; if the

statistician were computing health-insurance premiums the difference between .96 and .82 would be very important. In any such situation, the statistician must seek better data, that is, data rounded to the same degree of refinement.

The statistician often translates absolute data into percents and finds results somewhat as follows:

<i>Item</i>	<i>Budget Ratio, %</i>
Food.....	32
Rent.....	26
Clothing.....	12
Transportation.....	9
Medical care.....	6
Entertainment.....	4
Other.....	12
Total.....	101

The discrepancy of 1% arises from rounding; the statistician has rounded his results to the nearest 1%. But if he had rounded them to the nearest .1% he might have found:

<i>Item</i>	<i>Budget Ratio, %</i>
Food.....	31.6
Rent.....	25.6
Clothing.....	11.9
Transportation.....	8.5
Medical care.....	5.7
Entertainment.....	4.4
Other.....	12.3
Total.....	100.0

Refinement of figures will not necessarily or always abolish such a discrepancy. Moreover, such a refinement may not be permissible. The data from which the percentages are calculated are often not exact enough to justify extending a computation to another decimal place, and doing so would be a misuse of statistics. In such cases, however, the discrepancy between 100% and 101% is tolerable; if 32% is a satisfactory approximation for what may be any figure between 31.6% and 32.4%, then 101% is a satisfactory approximation of 100%. What the statistician must remember is that this 101% is

approximate, and cannot be treated in the same way as some other 101% that is not approximate.

Most people are familiar with the procedure of rounding approximate numbers:

1. If we drop a digit 6, 7, 8, or 9, we increase by 1 the preceding digit; thus 6.56 becomes 6.6 by rounding.

2. If we drop a digit 0, 1, 2, 3, or 4, we do not change the preceding digit; thus 6.54 becomes 6.5 by rounding.

3. If we drop a digit 5, we can usually assume that this digit would be followed (if the computation were carried further) by a digit 1 or greater in most cases and hence that it represents more than 5; on this assumption we increase the preceding digit by 1, and thus 6.55 becomes 6.6 by rounding. But it may be that the digit following the 5 is 0 and that the computation cannot validly be extended further. As a rule, in such a case we increase the preceding digit by 1 if it is odd, but leave it unchanged if it is even. Thus 16.550 becomes 16.6 by rounding and 16.650 also becomes 16.6 by rounding. In about half of all cases the effect is to make the rounded figure larger than the unrounded figure (as 16.6 is larger than 16.550); in the other half of all cases the effect is to make the rounded figure less than the unrounded figure (as 16.6 is less than 16.650). This treatment of the 50 digits, in rounding the figures of a long series that are to be added, avoids the cumulative errors that would result if all the 50's were treated in the same way.

III

III

$$\overline{5} \quad \overline{36.28}$$
$$\frac{2}{5}, \frac{1}{36}, \frac{1}{28}$$
$$\begin{array}{r} 2 \\ 5 \overline{) 36.28} \\ \underline{4} \\ 1 36 \end{array}$$

Then we double the result thus far achieved, that is, 2, and bring the 4 down as a "trial divisor" for 136. We disregard temporarily the last digit in the remainder—here 6—and divide 13 by 4. This gives a trial quotient of 3. We attach 3 to the trial divisor of 4, which gives us a new trial divisor of 43. We add the 3 to the quotient, above 36. Next we multiply the new trial divisor 43 by the 3 added to the quotient, and obtain 129. We subtract 129 from 136.

$$\begin{array}{r}
 2 \quad 3 \quad \_ \\
 5 \quad 36.28 \\
 4 \quad \_ \\
 43 \overline{) 1 \quad 36} \\
 \underline{1 \quad 29} \\
 7
 \end{array}$$

We place the decimal point in the result above the decimal point in the given number, that is, after the new quotient of 23, and bring down 28 to make a new remainder of 728.

$$\begin{array}{r}
 2 \quad 3 \quad \_ \\
 5 \quad 36.28 \\
 4 \quad \_ \\
 43 \overline{) 1 \quad 36} \\
 \underline{1 \quad 29} \\
 7 \quad 28
 \end{array}$$

Once again, we double the result thus far obtained—the 23—and bring the 46 down as a trial divisor for 728. Disregarding the last digit 8 in the remainder, we divide 72 by 46. This operation gives a trial quotient of 1, and we attach the 1 to the trial divisor of 46, which gives us a new trial divisor of 461. We add the 1 to the quotient, above 28. Next we multiply the trial divisor 461 by the 1 added to the quotient, and obtain 461. We subtract this 461 from 728.

$$\begin{array}{r}
 2 \quad 3 \quad 1 \quad \_ \\
 5 \quad 36.28 \\
 4 \quad \_ \\
 43 \overline{) 1 \quad 36} \\
 \underline{1 \quad 29} \\
 461 \overline{) 7 \quad 28} \\
 \underline{4 \quad 61} \\
 2 \quad 67
 \end{array}$$

To carry the answer out to additional decimal places, we add a *pair* (or pairs) of zeros to the given number 536.28 and bring down one pair of zeros at a time; proceeding as before, we arrive at:

$$\begin{array}{r}
 \frac{2}{5} \quad \frac{3}{36} \quad \frac{1}{28} \quad \frac{5}{00} \\
 4 \\
 \underline{43} \overline{) 1 \ 36} \\
 1 \ 29 \\
 \underline{461} \overline{) 7 \ 28} \\
 4 \ 61 \\
 \underline{4625} \overline{) 2 \ 67 \ 00} \\
 2 \ 31 \ 25 \\
 35 \ 75
 \end{array}$$

For additional decimal places, carry down additional pairs of zeros and proceed as above.

Appendix

IV

Table of Squares
and Square Roots

From *Rinehart Mathematical Tables*, Rinehart & Company, New York; copyright, 1948, by Harold D. Larsen.

1 — Squares and Square Roots — 100

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
1	1	1.00 000	3.16 228
2	4	1.41 421	4.47 214
3	9	1.73 205	5.47 723
4	16	2.00 000	6.32 456
5	25	2.23 607	7.07 107
6	36	2.44 949	7.74 597
7	49	2.64 575	8.36 660
8	64	2.82 843	8.94 427
9	81	3.00 000	9.48 683
10	100	3.16 228	10.00 00
11	121	3.31 662	10.48 81
12	144	3.46 410	10.95 45
13	169	3.60 555	11.40 18
14	196	3.74 166	11.83 22
15	225	3.87 298	12.24 74
16	256	4.00 000	12.64 91
17	289	4.12 311	13.03 84
18	324	4.24 264	13.41 64
19	361	4.35 890	13.78 40
20	400	4.47 214	14.14 21
21	441	4.58 258	14.49 14
22	484	4.69 042	14.83 24
23	529	4.79 583	15.16 58
24	576	4.89 898	15.49 19
25	625	5.00 000	15.81 14
26	676	5.09 902	16.12 45
27	729	5.19 615	16.43 17
28	784	5.29 150	16.73 32
29	841	5.38 516	17.02 94
30	900	5.47 723	17.32 05
31	961	5.56 776	17.60 68
32	1 024	5.65 685	17.88 85
33	1 089	5.74 456	18.16 59
34	1 156	5.83 095	18.43 91
35	1 225	5.91 608	18.70 83
36	1 296	6.00 000	18.97 37
37	1 369	6.08 276	19.23 54
38	1 444	6.16 441	19.49 36
39	1 521	6.24 500	19.74 84
40	1 600	6.32 456	20.00 00
41	1 681	6.40 312	20.24 85
42	1 764	6.48 074	20.49 39
43	1 849	6.55 744	20.73 64
44	1 936	6.63 325	20.97 62
45	2 025	6.70 820	21.21 32
46	2 116	6.78 233	21.44 76
47	2 209	6.85 565	21.67 95
48	2 304	6.92 820	21.90 89
49	2 401	7.00 000	22.13 59
50	2 500	7.07 107	22.36 07
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
50	2 500	7.07 107	22.36 07
51	2 601	7.14 143	22.58 32
52	2 704	7.21 110	22.80 35
53	2 809	7.28 011	23.02 17
54	2 916	7.34 847	23.23 79
55	3 025	7.41 620	23.45 21
56	3 136	7.48 331	23.66 43
57	3 249	7.54 983	23.87 47
58	3 364	7.61 577	24.08 32
59	3 481	7.68 115	24.28 99
60	3 600	7.74 597	24.49 49
61	3 721	7.81 025	24.69 82
62	3 844	7.87 401	24.89 98
63	3 969	7.93 725	25.09 98
64	4 096	8.00 000	25.29 82
65	4 225	8.06 226	25.49 51
66	4 356	8.12 404	25.69 05
67	4 489	8.18 535	25.88 44
68	4 624	8.24 621	26.07 68
69	4 761	8.30 662	26.26 79
70	4 900	8.36 660	26.45 76
71	5 041	8.42 615	26.64 58
72	5 184	8.48 528	26.83 28
73	5 329	8.54 400	27.01 85
74	5 476	8.60 233	27.20 29
75	5 625	8.66 025	27.38 61
76	5 776	8.71 780	27.56 81
77	5 929	8.77 496	27.74 89
78	6 084	8.83 176	27.92 85
79	6 241	8.88 819	28.10 69
80	6 400	8.94 427	28.28 43
81	6 561	9.00 000	28.46 05
82	6 724	9.05 539	28.63 56
83	6 889	9.11 043	28.80 97
84	7 056	9.16 515	28.98 28
85	7 225	9.21 954	29.15 48
86	7 396	9.27 362	29.32 58
87	7 569	9.32 738	29.49 58
88	7 744	9.38 083	29.66 48
89	7 921	9.43 398	29.83 29
90	8 100	9.48 683	30.00 00
91	8 281	9.53 939	30.16 62
92	8 464	9.59 166	30.33 15
93	8 649	9.64 365	30.49 59
94	8 836	9.69 536	30.65 94
95	9 025	9.74 679	30.82 21
96	9 216	9.79 796	30.98 39
97	9 409	9.84 886	31.14 48
98	9 604	9.89 949	31.30 50
99	9 801	9.94 987	31.46 43
100	10 000	10.00 000	31.62 28
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

100 — Squares and Square Roots — 200

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
100	10 000	10.00 00	31.62 28
101	10 201	10.04 99	31.78 05
102	10 404	10.09 95	31.93 74
103	10 609	10.14 89	32.09 36
104	10 816	10.19 80	32.24 90
105	11 025	10.24 70	32.40 37
106	11 236	10.29 56	32.55 76
107	11 449	10.34 41	32.71 09
108	11 664	10.39 23	32.86 34
109	11 881	10.44 03	33.01 51
110	12 100	10.48 81	33.16 62
111	12 321	10.53 57	33.31 67
112	12 544	10.58 30	33.46 64
113	12 769	10.63 01	33.61 55
114	12 996	10.67 71	33.76 39
115	13 225	10.72 38	33.91 16
116	13 456	10.77 03	34.05 88
117	13 689	10.81 67	34.20 53
118	13 924	10.86 28	34.35 11
119	14 161	10.90 87	34.49 64
120	14 400	10.95 45	34.64 10
121	14 641	11.00 00	34.78 51
122	14 884	11.04 54	34.92 85
123	15 129	11.09 05	35.07 14
124	15 376	11.13 55	35.21 36
125	15 625	11.18 03	35.35 53
126	15 876	11.22 50	35.49 65
127	16 129	11.26 94	35.63 71
128	16 384	11.31 37	35.77 71
129	16 641	11.35 78	35.91 66
130	16 900	11.40 18	36.05 55
131	17 161	11.44 55	36.19 39
132	17 424	11.48 91	36.33 18
133	17 689	11.53 26	36.46 92
134	17 956	11.57 58	36.60 60
135	18 225	11.61 90	36.74 23
136	18 496	11.66 19	36.87 82
137	18 769	11.70 47	37.01 35
138	19 044	11.74 73	37.14 84
139	19 321	11.78 98	37.28 27
140	19 600	11.83 22	37.41 66
141	19 881	11.87 43	37.55 00
142	20 164	11.91 64	37.68 29
143	20 449	11.95 83	37.81 53
144	20 736	12.00 00	37.94 73
145	21 025	12.04 16	38.07 89
146	21 316	12.08 30	38.20 99
147	21 609	12.12 44	38.34 06
148	21 904	12.16 55	38.47 08
149	22 201	12.20 66	38.60 05
150	22 500	12.24 74	38.72 98
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
150	22 500	12.24 74	38.72 98
151	22 801	12.28 82	38.85 87
152	23 104	12.32 88	38.98 72
153	23 409	12.36 93	39.11 52
154	23 716	12.40 97	39.24 28
155	24 025	12.44 99	39.37 00
156	24 336	12.49 00	39.49 68
157	24 649	12.53 00	39.62 32
158	24 964	12.56 98	39.74 92
159	25 281	12.60 95	39.87 48
160	25 600	12.64 91	40.00 00
161	25 921	12.68 86	40.12 48
162	26 244	12.72 79	40.24 92
163	26 569	12.76 71	40.37 33
164	26 896	12.80 62	40.49 69
165	27 225	12.84 52	40.62 02
166	27 556	12.88 41	40.74 31
167	27 889	12.92 28	40.86 56
168	28 224	12.96 15	40.98 78
169	28 561	13.00 00	41.10 96
170	28 900	13.03 84	41.23 11
171	29 241	13.07 67	41.35 21
172	29 584	13.11 49	41.47 29
173	29 929	13.15 29	41.59 33
174	30 276	13.19 09	41.71 33
175	30 625	13.22 88	41.83 30
176	30 976	13.26 65	41.95 24
177	31 329	13.30 41	42.07 14
178	31 684	13.34 17	42.19 00
179	32 041	13.37 91	42.30 84
180	32 400	13.41 64	42.42 64
181	32 761	13.45 36	42.54 41
182	33 124	13.49 07	42.66 15
183	33 489	13.52 77	42.77 85
184	33 856	13.56 47	42.89 52
185	34 225	13.60 15	43.01 16
186	34 596	13.63 82	43.12 77
187	34 969	13.67 48	43.24 35
188	35 344	13.71 13	43.35 90
189	35 721	13.74 77	43.47 41
190	36 100	13.78 40	43.58 90
191	36 481	13.82 03	43.70 35
192	36 864	13.85 64	43.81 78
193	37 249	13.89 24	43.93 18
194	37 636	13.92 84	44.04 54
195	38 025	13.96 42	44.15 88
196	38 416	14.00 00	44.27 19
197	38 809	14.03 57	44.38 47
198	39 204	14.07 12	44.49 72
199	39 601	14.10 67	44.60 94
200	40 000	14.14 21	44.72 14
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

200 — Squares and Square Roots — 300

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
200	40 000	14.14 21	44.72 14
201	40 401	14.17 74	44.83 30
202	40 804	14.21 27	44.94 44
203	41 209	14.24 78	45.05 55
204	41 616	14.28 29	45.16 64
205	42 025	14.31 78	45.27 69
206	42 436	14.35 27	45.38 72
207	42 849	14.38 75	45.49 73
208	43 264	14.42 22	45.60 70
209	43 681	14.45 68	45.71 65
210	44 100	14.49 14	45.82 58
211	44 521	14.52 58	45.93 47
212	44 944	14.56 02	46.04 35
213	45 369	14.59 45	46.15 19
214	45 796	14.62 87	46.26 01
215	46 225	14.66 29	46.36 81
216	46 656	14.69 69	46.47 58
217	47 089	14.73 09	46.58 33
218	47 524	14.76 48	46.69 05
219	47 961	14.79 86	46.79 74
220	48 400	14.83 24	46.90 42
221	48 841	14.86 61	47.01 06
222	49 284	14.89 97	47.11 69
223	49 729	14.93 32	47.22 29
224	50 176	14.96 66	47.32 86
225	50 625	15.00 00	47.43 42
226	51 076	15.03 33	47.53 95
227	51 529	15.06 65	47.64 45
228	51 984	15.09 97	47.74 93
229	52 441	15.13 27	47.85 39
230	52 900	15.16 58	47.95 83
231	53 361	15.19 87	48.06 25
232	53 824	15.23 15	48.16 64
233	54 289	15.26 43	48.27 01
234	54 756	15.29 71	48.37 35
235	55 225	15.32 97	48.47 68
236	55 696	15.36 23	48.57 98
237	56 169	15.39 48	48.68 26
238	56 644	15.42 72	48.78 52
239	57 121	15.45 96	48.88 76
240	57 600	15.49 19	48.98 98
241	58 081	15.52 42	49.09 18
242	58 564	15.55 63	49.19 35
243	59 049	15.58 85	49.29 50
244	59 536	15.62 05	49.39 64
245	60 025	15.65 25	49.49 75
246	60 516	15.68 44	49.59 84
247	61 009	15.71 62	49.69 91
248	61 504	15.74 80	49.79 96
249	62 001	15.77 97	49.89 99
250	62 500	15.81 14	50.00 00
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
250	62 500	15.81 14	50.00 00
251	63 001	15.84 30	50.09 99
252	63 504	15.87 45	50.19 96
253	64 009	15.90 60	50.29 91
254	64 516	15.93 74	50.39 84
255	65 025	15.96 87	50.49 75
256	65 536	16.00 00	50.59 64
257	66 049	16.03 12	50.69 62
258	66 564	16.06 24	50.79 57
259	67 081	16.09 35	50.89 20
260	67 600	16.12 45	50.99 02
261	68 121	16.15 55	51.08 82
262	68 644	16.18 64	51.18 59
263	69 169	16.21 73	51.28 35
264	69 696	16.24 81	51.38 09
265	70 225	16.27 88	51.47 82
266	70 756	16.30 95	51.57 52
267	71 289	16.34 01	51.67 20
268	71 824	16.37 07	51.76 87
269	72 361	16.40 12	51.86 52
270	72 900	16.43 17	51.96 16
271	73 441	16.46 21	52.05 77
272	73 984	16.49 24	52.15 36
273	74 529	16.52 27	52.24 94
274	75 076	16.55 29	52.34 50
275	75 625	16.58 31	52.44 04
276	76 176	16.61 32	52.53 57
277	76 729	16.64 33	52.63 00
278	77 284	16.67 33	52.72 57
279	77 841	16.70 33	52.82 05
280	78 400	16.73 32	52.91 50
281	78 961	16.76 31	53.00 94
282	79 524	16.79 29	53.10 37
283	80 089	16.82 26	53.19 77
284	80 656	16.85 23	53.29 17
285	81 225	16.88 19	53.38 54
286	81 796	16.91 15	53.47 90
287	82 369	16.94 11	53.57 24
288	82 944	16.97 06	53.66 56
289	83 521	17.00 00	53.75 87
290	84 100	17.02 94	53.85 16
291	84 681	17.05 87	53.94 44
292	85 264	17.08 80	54.03 70
293	85 849	17.11 72	54.12 95
294	86 436	17.14 64	54.22 18
295	87 025	17.17 56	54.31 39
296	87 616	17.20 47	54.40 59
297	88 209	17.23 37	54.49 77
298	88 804	17.26 27	54.58 94
299	89 401	17.29 16	54.68 09
300	90 000	17.32 05	54.77 23
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

300 — Squares and Square Roots — 400

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
300	90 000	17.32 05	54.77 23
301	90 601	17.34 94	54.86 35
302	91 204	17.37 81	54.95 45
303	91 809	17.40 69	55.04 54
304	92 416	17.43 56	55.13 62
305	93 025	17.46 42	55.22 68
306	93 636	17.49 29	55.31 73
307	94 249	17.52 14	55.40 76
308	94 864	17.54 99	55.49 77
309	95 481	17.57 84	55.58 78
310	96 100	17.60 68	55.67 76
311	96 721	17.63 52	55.76 74
312	97 344	17.66 35	55.85 70
313	97 969	17.69 18	55.94 64
314	98 596	17.72 00	56.03 57
315	99 225	17.74 82	56.12 49
316	99 856	17.77 64	56.21 39
317	100 489	17.80 45	56.30 28
318	101 124	17.83 26	56.39 15
319	101 761	17.86 06	56.48 01
320	102 400	17.88 85	56.56 85
321	103 041	17.91 65	56.65 69
322	103 684	17.94 44	56.74 50
323	104 329	17.97 22	56.83 31
324	104 976	18.00 00	56.92 10
325	105 625	18.02 78	57.00 88
326	106 276	18.05 55	57.09 64
327	106 929	18.08 31	57.18 39
328	107 584	18.11 08	57.27 13
329	108 241	18.13 84	57.35 85
330	108 900	18.16 59	57.44 56
331	109 561	18.19 34	57.53 26
332	110 224	18.22 09	57.61 94
333	110 889	18.24 83	57.70 62
334	111 556	18.27 57	57.79 27
335	112 225	18.30 30	57.87 92
336	112 896	18.33 03	57.96 55
337	113 569	18.35 76	58.05 17
338	114 244	18.38 48	58.13 78
339	114 921	18.41 20	58.22 37
340	115 600	18.43 91	58.30 95
341	116 281	18.46 62	58.39 52
342	116 964	18.49 32	58.48 08
343	117 649	18.52 03	58.56 62
344	118 336	18.54 72	58.65 15
345	119 025	18.57 42	58.73 67
346	119 716	18.60 11	58.82 18
347	120 409	18.62 79	58.90 67
348	121 104	18.65 48	58.99 15
349	121 801	18.68 15	59.07 62
350	122 500	18.70 83	59.16 08
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
350	122 500	18.70 83	59.16 08
351	123 201	18.73 50	59.24 53
352	123 904	18.76 17	59.32 96
353	124 609	18.78 83	59.41 38
354	125 316	18.81 49	59.49 79
355	126 025	18.84 14	59.58 19
356	126 736	18.86 80	59.66 57
357	127 449	18.89 44	59.74 95
358	128 164	18.92 09	59.83 31
359	128 881	18.94 73	59.91 66
360	129 600	18.97 37	60.00 00
361	130 321	19.00 00	60.08 33
362	131 044	19.02 63	60.16 64
363	131 769	19.05 26	60.24 95
364	132 496	19.07 88	60.33 24
365	133 225	19.10 50	60.41 52
366	133 956	19.13 11	60.49 79
367	134 689	19.15 72	60.58 05
368	135 424	19.18 33	60.66 30
369	136 161	19.20 94	60.74 54
370	136 900	19.23 54	60.82 76
371	137 641	19.26 14	60.90 98
372	138 384	19.28 73	60.99 18
373	139 129	19.31 32	61.07 37
374	139 876	19.33 91	61.15 55
375	140 625	19.36 49	61.23 72
376	141 376	19.39 07	61.31 88
377	142 129	19.41 65	61.40 03
378	142 884	19.44 22	61.48 17
379	143 641	19.46 79	61.56 30
380	144 400	19.49 36	61.64 41
381	145 161	19.51 92	61.72 52
382	145 924	19.54 48	61.80 61
383	146 689	19.57 04	61.88 70
384	147 456	19.59 59	61.96 77
385	148 225	19.62 14	62.04 84
386	148 996	19.64 69	62.12 89
387	149 769	19.67 23	62.20 93
388	150 544	19.69 77	62.28 96
389	151 321	19.72 31	62.36 99
390	152 100	19.74 84	62.45 00
391	152 881	19.77 37	62.53 00
392	153 664	19.79 90	62.60 99
393	154 449	19.82 42	62.68 97
394	155 236	19.84 94	62.76 94
395	156 025	19.87 46	62.84 90
396	156 816	19.89 97	62.92 85
397	157 609	19.92 49	63.00 79
398	158 404	19.94 99	63.08 72
399	159 201	19.97 50	63.16 64
400	160 000	20.00 00	63.24 56
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

400 — Squares and Square Roots — 500

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
400	160 000	20.00 00	63.24 56
401	160 801	20.02 50	63.32 46
402	161 604	20.04 99	63.40 35
403	162 409	20.07 49	63.48 23
404	163 216	20.09 98	63.56 10
405	164 025	20.12 46	63.63 96
406	164 836	20.14 94	63.71 81
407	165 649	20.17 42	63.79 66
408	166 464	20.19 90	63.87 49
409	167 281	20.22 37	63.95 31
410	168 100	20.24 85	64.03 12
411	168 921	20.27 31	64.10 93
412	169 744	20.29 78	64.18 72
413	170 569	20.32 24	64.26 51
414	171 396	20.34 70	64.34 28
415	172 225	20.37 15	64.42 05
416	173 056	20.39 61	64.49 81
417	173 889	20.42 06	64.57 55
418	174 724	20.44 50	64.65 29
419	175 561	20.46 95	64.73 02
420	176 400	20.49 39	64.80 74
421	177 241	20.51 83	64.88 45
422	178 084	20.54 26	64.96 15
423	178 929	20.56 70	65.03 85
424	179 776	20.59 13	65.11 53
425	180 625	20.61 55	65.19 20
426	181 476	20.63 98	65.26 87
427	182 329	20.66 40	65.34 52
428	183 184	20.68 82	65.42 17
429	184 041	20.71 23	65.49 81
430	184 900	20.73 64	65.57 44
431	185 761	20.76 05	65.65 06
432	186 624	20.78 46	65.72 67
433	187 489	20.80 87	65.80 27
434	188 356	20.83 27	65.87 87
435	189 225	20.85 67	65.95 45
436	190 096	20.88 06	66.03 03
437	190 969	20.90 45	66.10 60
438	191 844	20.92 84	66.18 16
439	192 721	20.95 23	66.25 71
440	193 600	20.97 62	66.33 25
441	194 481	21.00 00	66.40 78
442	195 364	21.02 38	66.48 31
443	196 249	21.04 76	66.55 82
444	197 136	21.07 13	66.63 33
445	198 025	21.09 50	66.70 83
446	198 916	21.11 87	66.78 32
447	199 809	21.14 24	66.85 81
448	200 704	21.16 60	66.93 28
449	201 601	21.18 96	67.00 75
450	202 500	21.21 32	67.08 20
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
450	202 500	21.21 32	67.08 20
451	203 401	21.23 68	67.15 65
452	204 304	21.26 03	67.23 09
453	205 209	21.28 38	67.30 53
454	206 116	21.30 73	67.37 95
455	207 025	21.33 07	67.45 37
456	207 936	21.35 42	67.52 78
457	208 849	21.37 76	67.60 18
458	209 764	21.40 09	67.67 57
459	210 681	21.42 43	67.74 95
460	211 600	21.44 76	67.82 33
461	212 521	21.47 09	67.89 70
462	213 444	21.49 42	67.97 06
463	214 369	21.51 74	68.04 41
464	215 296	21.54 07	68.11 75
465	216 225	21.56 39	68.19 09
466	217 156	21.58 70	68.26 42
467	218 089	21.61 02	68.33 74
468	219 024	21.63 33	68.41 05
469	219 961	21.65 64	68.48 36
470	220 900	21.67 95	68.55 65
471	221 841	21.70 25	68.62 94
472	222 784	21.72 56	68.70 23
473	223 729	21.74 86	68.77 50
474	224 676	21.77 15	68.84 77
475	225 625	21.79 45	68.92 02
476	226 576	21.81 74	68.99 28
477	227 529	21.84 03	69.06 52
478	228 484	21.86 32	69.13 75
479	229 441	21.88 61	69.20 98
480	230 400	21.90 89	69.28 20
481	231 361	21.93 17	69.35 42
482	232 324	21.95 45	69.42 62
483	233 289	21.97 73	69.49 82
484	234 256	22.00 00	69.57 01
485	235 225	22.02 27	69.64 19
486	236 196	22.04 54	69.71 37
487	237 169	22.06 81	69.78 54
488	238 144	22.09 07	69.85 70
489	239 121	22.11 33	69.92 85
490	240 100	22.13 59	70.00 00
491	241 081	22.15 85	70.07 14
492	242 064	22.18 11	70.14 27
493	243 049	22.20 36	70.21 40
494	244 036	22.22 61	70.28 51
495	245 025	22.24 86	70.35 62
496	246 016	22.27 11	70.42 73
497	247 009	22.29 35	70.49 82
498	248 004	22.31 59	70.56 91
499	249 001	22.33 83	70.63 99
500	250 000	22.36 07	70.71 07
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

500 — Squares and Square Roots — 600

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
500	250 000	22.36 07	70.71 07
501	251 001	22.38 30	70.78 14
502	252 004	22.40 54	70.85 20
503	253 009	22.42 77	70.92 25
504	254 016	22.44 99	70.99 30
505	255 025	22.47 22	71.06 34
506	256 036	22.49 44	71.13 37
507	257 049	22.51 67	71.20 39
508	258 064	22.53 89	71.27 41
509	259 081	22.56 10	71.34 42
510	260 100	22.58 32	71.41 43
511	261 121	22.60 53	71.48 45
512	262 144	22.62 74	71.55 46
513	263 169	22.64 95	71.62 47
514	264 196	22.67 16	71.69 48
515	265 225	22.69 36	71.76 49
516	266 256	22.71 56	71.83 51
517	267 289	22.73 76	71.90 52
518	268 324	22.75 96	71.97 53
519	269 361	22.78 16	72.04 54
520	270 400	22.80 35	72.11 55
521	271 441	22.82 54	72.18 56
522	272 484	22.84 73	72.25 57
523	273 529	22.86 92	72.31 58
524	274 576	22.89 10	72.38 59
525	275 625	22.91 29	72.45 60
526	276 676	22.93 47	72.52 61
527	277 729	22.95 65	72.59 62
528	278 784	22.97 83	72.66 63
529	279 841	23.00 00	72.73 64
530	280 900	23.02 17	72.80 65
531	281 961	23.04 34	72.86 67
532	283 024	23.06 51	72.93 68
533	284 089	23.08 68	73.00 69
534	285 156	23.10 84	73.07 70
535	286 225	23.13 01	73.14 71
536	287 296	23.15 17	73.21 72
537	288 369	23.17 33	73.28 73
538	289 444	23.19 48	73.34 74
539	290 521	23.21 64	73.41 75
540	291 600	23.23 79	73.48 76
541	292 681	23.25 94	73.55 77
542	293 764	23.28 09	73.62 78
543	294 849	23.30 24	73.68 79
544	295 936	23.32 38	73.75 80
545	297 025	23.34 52	73.82 81
546	298 116	23.36 66	73.89 82
547	299 209	23.38 80	73.95 83
548	300 304	23.40 94	74.02 84
549	301 401	23.43 07	74.09 85
550	302 500	23.45 21	74.16 86
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
550	302 500	23.45 21	74.16 20
551	303 601	23.47 34	74.22 94
552	304 704	23.49 47	74.29 67
553	305 809	23.51 60	74.36 40
554	306 916	23.53 72	74.43 12
555	308 025	23.55 84	74.49 83
556	309 136	23.57 97	74.56 54
557	310 249	23.60 08	74.63 24
558	311 364	23.62 20	74.69 94
559	312 481	23.64 32	74.76 63
560	313 600	23.66 43	74.83 31
561	314 721	23.68 54	74.89 99
562	315 844	23.70 65	74.96 67
563	316 969	23.72 76	75.03 33
564	318 096	23.74 87	75.09 99
565	319 225	23.76 97	75.16 65
566	320 356	23.79 08	75.23 30
567	321 489	23.81 18	75.29 94
568	322 624	23.83 28	75.36 58
569	323 761	23.85 37	75.43 21
570	324 900	23.87 47	75.49 83
571	326 041	23.89 56	75.56 45
572	327 184	23.91 65	75.63 07
573	328 329	23.93 74	75.69 68
574	329 476	23.95 83	75.76 28
575	330 625	23.97 92	75.82 88
576	331 776	24.00 00	75.89 47
577	332 929	24.02 08	75.96 05
578	334 084	24.04 16	76.02 63
579	335 241	24.06 24	76.09 20
580	336 400	24.08 32	76.15 77
581	337 561	24.10 39	76.22 34
582	338 724	24.12 47	76.28 89
583	339 889	24.14 54	76.35 44
584	341 056	24.16 61	76.41 99
585	342 225	24.18 68	76.48 53
586	343 396	24.20 74	76.55 06
587	344 569	24.22 81	76.61 59
588	345 744	24.24 87	76.68 12
589	346 921	24.26 93	76.74 63
590	348 100	24.28 99	76.81 15
591	349 281	24.31 05	76.87 65
592	350 464	24.33 11	76.94 15
593	351 649	24.35 16	77.00 65
594	352 836	24.37 21	77.07 14
595	354 025	24.39 26	77.13 62
596	355 216	24.41 31	77.20 10
597	356 409	24.43 36	77.26 58
598	357 604	24.45 40	77.33 05
599	358 801	24.47 45	77.39 51
600	360 000	24.49 49	77.45 97
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

600 — Squares and Square Roots — 700

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
600	360 000	24.49 49	77.45 97
601	361 201	24.51 53	77.52 42
602	362 404	24.53 57	77.58 87
603	363 609	24.55 61	77.65 31
604	364 816	24.57 64	77.71 74
605	366 025	24.59 67	77.78 17
606	367 236	24.61 71	77.84 60
607	368 449	24.63 74	77.91 02
608	369 664	24.65 77	77.97 44
609	370 881	24.67 79	78.03 85
610	372 100	24.69 82	78.10 25
611	373 321	24.71 84	78.16 65
612	374 544	24.73 86	78.23 04
613	375 769	24.75 88	78.29 43
614	376 996	24.77 90	78.35 82
615	378 225	24.79 92	78.42 19
616	379 456	24.81 93	78.48 57
617	380 689	24.83 95	78.54 93
618	381 924	24.85 96	78.61 30
619	383 161	24.87 97	78.67 66
620	384 400	24.89 98	78.74 01
621	385 641	24.91 99	78.80 36
622	386 884	24.93 99	78.86 70
623	388 129	24.96 00	78.93 03
624	389 376	24.98 00	78.99 37
625	390 625	25.00 00	79.05 69
626	391 876	25.02 00	79.12 02
627	393 129	25.04 00	79.18 33
628	394 384	25.05 99	79.24 65
629	395 641	25.07 99	79.30 95
630	396 900	25.09 98	79.37 25
631	398 161	25.11 97	79.43 55
632	399 424	25.13 96	79.49 84
633	400 689	25.15 95	79.56 13
634	401 956	25.17 94	79.62 41
635	403 225	25.19 92	79.68 69
636	404 496	25.21 90	79.74 96
637	405 769	25.23 89	79.81 23
638	407 044	25.25 87	79.87 49
639	408 321	25.27 84	79.93 75
640	409 600	25.29 82	80.00 00
641	410 881	25.31 80	80.06 25
642	412 164	25.33 77	80.12 49
643	413 449	25.35 74	80.18 73
644	414 736	25.37 72	80.24 96
645	416 025	25.39 69	80.31 19
646	417 316	25.41 65	80.37 41
647	418 609	25.43 62	80.43 63
648	419 904	25.45 58	80.49 84
649	421 201	25.47 55	80.56 05
650	422 500	25.49 51	80.62 26
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
650	422 500	25.49 51	80.62 26
651	423 801	25.51 47	80.68 46
652	425 104	25.53 43	80.74 65
653	426 409	25.55 39	80.80 84
654	427 716	25.57 34	80.87 03
655	429 025	25.59 30	80.93 21
656	430 336	25.61 25	80.99 38
657	431 649	25.63 20	81.05 55
658	432 964	25.65 15	81.11 72
659	434 281	25.67 10	81.17 88
660	435 600	25.69 05	81.24 04
661	436 921	25.70 99	81.30 19
662	438 244	25.72 94	81.36 34
663	439 569	25.74 88	81.42 48
664	440 896	25.76 82	81.48 62
665	442 225	25.78 76	81.54 75
666	443 556	25.80 70	81.60 88
667	444 889	25.82 63	81.67 01
668	446 224	25.84 57	81.73 13
669	447 561	25.86 50	81.79 24
670	448 900	25.88 44	81.85 35
671	450 241	25.90 37	81.91 46
672	451 584	25.92 30	81.97 56
673	452 929	25.94 22	82.03 66
674	454 276	25.96 15	82.09 75
675	455 625	25.98 08	82.15 84
676	456 976	26.00 00	82.21 92
677	458 329	26.01 92	82.28 00
678	459 684	26.03 84	82.34 08
679	461 041	26.05 76	82.40 15
680	462 400	26.07 68	82.46 21
681	463 761	26.09 60	82.52 27
682	465 124	26.11 51	82.58 33
683	466 489	26.13 43	82.64 38
684	467 856	26.15 34	82.70 43
685	469 225	26.17 25	82.76 47
686	470 596	26.19 16	82.82 51
687	471 969	26.21 07	82.88 55
688	473 344	26.22 98	82.94 58
689	474 721	26.24 88	83.00 60
690	476 100	26.26 79	83.06 62
691	477 481	26.28 69	83.12 64
692	478 864	26.30 59	83.18 65
693	480 249	26.32 49	83.24 66
694	481 636	26.34 39	83.30 67
695	483 025	26.36 29	83.36 67
696	484 416	26.38 18	83.42 66
697	485 809	26.40 08	83.48 65
698	487 204	26.41 97	83.54 64
699	488 601	26.43 86	83.60 62
700	490 000	26.45 75	83.66 60
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

700 — Squares and Square Roots — 800

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
700	490 000	26.45 75	83.66 60
701	491 401	26.47 64	83.72 57
702	492 804	26.49 53	83.78 54
703	494 209	26.51 41	83.84 51
704	495 616	26.53 30	83.90 47
705	497 025	26.55 18	83.96 43
706	498 436	26.57 07	84.02 38
707	499 849	26.58 95	84.08 33
708	501 264	26.60 83	84.14 27
709	502 681	26.62 71	84.20 21
710	504 100	26.64 58	84.26 15
711	505 521	26.66 46	84.32 08
712	506 944	26.68 33	84.38 01
713	508 369	26.70 21	84.43 93
714	509 796	26.72 08	84.49 85
715	511 225	26.73 95	84.55 77
716	512 656	26.75 82	84.61 68
717	514 089	26.77 69	84.67 59
718	515 524	26.79 55	84.73 49
719	516 961	26.81 42	84.79 39
720	518 400	26.83 28	84.85 28
721	519 841	26.85 14	84.91 17
722	521 284	26.87 01	84.97 06
723	522 729	26.88 87	85.02 94
724	524 176	26.90 72	85.08 82
725	525 625	26.92 58	85.14 69
726	527 076	26.94 44	85.20 56
727	528 529	26.96 29	85.26 43
728	529 984	26.98 15	85.32 29
729	531 441	27.00 00	85.38 15
730	532 900	27.01 85	85.44 00
731	534 361	27.03 70	85.49 85
732	535 824	27.05 55	85.55 70
733	537 289	27.07 40	85.61 54
734	538 756	27.09 24	85.67 38
735	540 225	27.11 09	85.73 21
736	541 696	27.12 93	85.79 04
737	543 169	27.14 77	85.84 87
738	544 644	27.16 62	85.90 69
739	546 121	27.18 46	85.96 51
740	547 600	27.20 29	86.02 33
741	549 081	27.22 13	86.08 14
742	550 564	27.23 97	86.13 94
743	552 049	27.25 80	86.19 74
744	553 536	27.27 64	86.25 54
745	555 025	27.29 47	86.31 34
746	556 516	27.31 30	86.37 13
747	558 009	27.33 13	86.42 92
748	559 504	27.34 96	86.48 70
749	561 001	27.36 79	86.54 48
750	562 500	27.38 61	86.60 25
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
750	562 500	27.38 61	86.60 25
751	564 001	27.40 44	86.66 03
752	565 504	27.42 26	86.71 79
753	567 009	27.44 08	86.77 56
754	568 516	27.45 91	86.83 32
755	570 025	27.47 73	86.89 07
756	571 536	27.49 55	86.94 83
757	573 049	27.51 36	87.00 57
758	574 564	27.53 18	87.06 32
759	576 081	27.55 00	87.12 06
760	577 600	27.56 81	87.17 80
761	579 121	27.58 62	87.23 53
762	580 644	27.60 43	87.29 26
763	582 169	27.62 25	87.34 99
764	583 696	27.64 05	87.40 71
765	585 225	27.65 86	87.46 43
766	586 756	27.67 67	87.52 14
767	588 289	27.69 48	87.57 85
768	589 824	27.71 28	87.63 56
769	591 361	27.73 08	87.69 26
770	592 900	27.74 89	87.74 96
771	594 441	27.76 69	87.80 66
772	595 984	27.78 49	87.86 35
773	597 529	27.80 29	87.92 04
774	599 076	27.82 09	87.97 73
775	600 625	27.83 88	88.03 41
776	602 176	27.85 68	88.09 09
777	603 729	27.87 47	88.14 76
778	605 284	27.89 27	88.20 43
779	606 841	27.91 06	88.26 10
780	608 400	27.92 85	88.31 76
781	609 961	27.94 64	88.37 42
782	611 524	27.96 43	88.43 08
783	613 089	27.98 21	88.48 73
784	614 656	28.00 00	88.54 38
785	616 225	28.01 79	88.60 02
786	617 796	28.03 57	88.65 66
787	619 369	28.05 35	88.71 30
788	620 944	28.07 13	88.76 94
789	622 521	28.08 91	88.82 57
790	624 100	28.10 69	88.88 19
791	625 681	28.12 47	88.93 82
792	627 264	28.14 25	88.99 44
793	628 849	28.16 03	89.05 05
794	630 436	28.17 80	89.10 67
795	632 025	28.19 57	89.16 28
796	633 616	28.21 35	89.21 88
797	635 209	28.23 12	89.27 49
798	636 804	28.24 89	89.33 08
799	638 401	28.26 66	89.38 68
800	640 000	28.28 43	89.44 27
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

800 — Squares and Square Roots — 900

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
800	640 000	28.28 43	89.44 27
801	641 601	28.30 19	89.49 86
802	643 204	28.31 96	89.55 45
803	644 809	28.33 73	89.61 03
804	646 416	28.35 49	89.66 60
805	648 025	28.37 25	89.72 18
806	649 636	28.39 01	89.77 75
807	651 249	28.40 77	89.83 32
808	652 864	28.42 53	89.88 88
809	654 481	28.44 29	89.94 44
810	656 100	28.46 05	90.00 00
811	657 721	28.47 81	90.05 55
812	659 344	28.49 56	90.11 10
813	660 969	28.51 32	90.16 65
814	662 596	28.53 07	90.22 19
815	664 225	28.54 82	90.27 74
816	665 856	28.56 57	90.33 27
817	667 489	28.58 32	90.38 81
818	669 124	28.60 07	90.44 34
819	670 761	28.61 82	90.49 86
820	672 400	28.63 56	90.55 39
821	674 041	28.65 31	90.60 91
822	675 684	28.67 05	90.66 42
823	677 329	28.68 80	90.71 93
824	678 976	28.70 54	90.77 44
825	680 625	28.72 28	90.82 95
826	682 276	28.74 02	90.88 45
827	683 929	28.75 76	90.93 95
828	685 584	28.77 50	90.99 45
829	687 241	28.79 24	91.04 94
830	688 900	28.80 97	91.10 43
831	690 561	28.82 71	91.15 92
832	692 224	28.84 44	91.21 40
833	693 889	28.86 17	91.26 88
834	695 556	28.87 91	91.32 36
835	697 225	28.89 64	91.37 83
836	698 896	28.91 37	91.43 30
837	700 569	28.93 10	91.48 77
838	702 244	28.94 82	91.54 23
839	703 921	28.96 55	91.59 69
840	705 600	28.98 28	91.65 15
841	707 281	29.00 00	91.70 61
842	708 964	29.01 72	91.76 06
843	710 649	29.03 45	91.81 50
844	712 336	29.05 17	91.86 95
845	714 025	29.06 89	91.92 39
846	715 716	29.08 61	91.97 83
847	717 409	29.10 33	92.03 26
848	719 104	29.12 04	92.08 69
849	720 801	29.13 76	92.14 12
850	722 500	29.15 48	92.19 54
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
850	722 500	29.15 48	92.19 54
851	724 201	29.17 19	92.24 97
852	725 904	29.18 90	92.30 38
853	727 609	29.20 62	92.35 80
854	729 316	29.22 33	92.41 21
855	731 025	29.24 04	92.46 62
856	732 736	29.25 75	92.52 03
857	734 449	29.27 46	92.57 43
858	736 164	29.29 16	92.62 83
859	737 881	29.30 87	92.68 23
860	739 600	29.32 58	92.73 62
861	741 321	29.34 28	92.79 01
862	743 044	29.35 98	92.84 40
863	744 769	29.37 69	92.89 78
864	746 496	29.39 39	92.95 16
865	748 225	29.41 09	93.00 54
866	749 956	29.42 79	93.05 91
867	751 689	29.44 49	93.11 28
868	753 424	29.46 18	93.16 65
869	755 161	29.47 88	93.22 02
870	756 900	29.49 58	93.27 38
871	758 641	29.51 27	93.32 74
872	760 384	29.52 96	93.38 09
873	762 129	29.54 66	93.43 45
874	763 876	29.56 35	93.48 80
875	765 625	29.58 04	93.54 14
876	767 376	29.59 73	93.59 49
877	769 129	29.61 42	93.64 83
878	770 884	29.63 11	93.70 17
879	772 641	29.64 79	93.75 50
880	774 400	29.66 48	93.80 83
881	776 161	29.68 16	93.86 16
882	777 924	29.69 85	93.91 49
883	779 689	29.71 53	93.96 81
884	781 456	29.73 21	94.02 13
885	783 225	29.74 89	94.07 44
886	784 996	29.76 58	94.12 76
887	786 769	29.78 25	94.18 07
888	788 544	29.79 93	94.23 38
889	790 321	29.81 61	94.28 68
890	792 100	29.83 29	94.33 98
891	793 881	29.84 96	94.39 28
892	795 664	29.86 64	94.44 58
893	797 449	29.88 31	94.49 87
894	799 236	29.89 98	94.55 16
895	801 025	29.91 66	94.60 44
896	802 816	29.93 33	94.65 73
897	804 609	29.95 00	94.71 01
898	806 404	29.96 66	94.76 29
899	808 201	29.98 33	94.81 56
900	810 000	30.00 00	94.86 83
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

900 — Squares and Square Roots — 1000

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
900	810 000	30.00 00	94.86 83
901	811 801	30.01 67	94.92 10
902	813 604	30.03 33	94.97 37
903	815 409	30.05 00	95.02 63
904	817 216	30.06 66	95.07 89
905	819 025	30.08 32	95.13 15
906	820 836	30.09 98	95.18 40
907	822 649	30.11 64	95.23 65
908	824 464	30.13 30	95.28 90
909	826 281	30.14 96	95.34 15
910	828 100	30.16 62	95.39 39
911	829 921	30.18 28	95.44 63
912	831 744	30.19 93	95.49 87
913	833 569	30.21 59	95.55 10
914	835 396	30.23 24	95.60 33
915	837 225	30.24 90	95.65 56
916	839 056	30.26 55	95.70 79
917	840 889	30.28 20	95.76 01
918	842 724	30.29 85	95.81 23
919	844 561	30.31 50	95.86 45
920	846 400	30.33 15	95.91 66
921	848 241	30.34 80	95.96 87
922	850 084	30.36 45	96.02 08
923	851 929	30.38 09	96.07 29
924	853 776	30.39 74	96.12 49
925	855 625	30.41 38	96.17 69
926	857 476	30.43 02	96.22 89
927	859 329	30.44 67	96.28 08
928	861 184	30.46 31	96.33 28
929	863 041	30.47 95	96.38 46
930	864 900	30.49 59	96.43 65
931	866 761	30.51 23	96.48 83
932	868 624	30.52 87	96.54 01
933	870 489	30.54 50	96.59 19
934	872 356	30.56 14	96.64 37
935	874 225	30.57 78	96.69 54
936	876 096	30.59 41	96.74 71
937	877 969	30.61 05	96.79 88
938	879 844	30.62 68	96.85 04
939	881 721	30.64 31	96.90 20
940	883 600	30.65 94	96.95 36
941	885 481	30.67 57	97.00 52
942	887 364	30.69 20	97.05 67
943	889 249	30.70 83	97.10 82
944	891 136	30.72 46	97.15 97
945	893 025	30.74 09	97.21 11
946	894 916	30.75 71	97.26 25
947	896 809	30.77 34	97.31 39
948	898 704	30.78 96	97.36 53
949	900 601	30.80 58	97.41 66
950	902 500	30.82 21	97.46 79
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
950	902 500	30.82 21	97.46 79
951	904 401	30.83 83	97.51 92
952	906 304	30.85 45	97.57 05
953	908 209	30.87 07	97.62 17
954	910 116	30.88 69	97.67 29
955	912 025	30.90 31	97.72 41
956	913 936	30.91 92	97.77 53
957	915 849	30.93 54	97.82 64
958	917 764	30.95 16	97.87 75
959	919 681	30.96 77	97.92 85
960	921 600	30.98 39	97.97 96
961	923 521	31.00 00	98.03 06
962	925 444	31.01 61	98.08 16
963	927 369	31.03 22	98.13 26
964	929 296	31.04 83	98.18 35
965	931 225	31.06 44	98.23 44
966	933 156	31.08 05	98.28 53
967	935 089	31.09 66	98.33 62
968	937 024	31.11 27	98.38 70
969	938 961	31.12 88	98.43 78
970	940 900	31.14 48	98.48 86
971	942 841	31.16 09	98.53 93
972	944 784	31.17 69	98.59 01
973	946 729	31.19 29	98.64 08
974	948 676	31.20 90	98.69 14
975	950 625	31.22 50	98.74 21
976	952 576	31.24 10	98.79 27
977	954 529	31.25 70	98.84 33
978	956 484	31.27 30	98.89 39
979	958 441	31.28 90	98.94 44
980	960 400	31.30 50	98.99 49
981	962 361	31.32 09	99.04 54
982	964 324	31.33 69	99.09 59
983	966 289	31.35 28	99.14 64
984	968 256	31.36 88	99.19 68
985	970 225	31.38 47	99.24 72
986	972 196	31.40 06	99.29 75
987	974 169	31.41 66	99.34 79
988	976 144	31.43 25	99.39 82
989	978 121	31.44 84	99.44 85
990	980 100	31.46 43	99.49 87
991	982 081	31.48 02	99.54 90
992	984 064	31.49 60	99.59 92
993	986 049	31.51 19	99.64 94
994	988 036	31.52 78	99.69 95
995	990 025	31.54 36	99.74 97
996	992 016	31.55 95	99.79 98
997	994 009	31.57 53	99.84 99
998	996 004	31.59 11	99.89 99
999	998 001	31.60 70	99.95 00
1000	1000 000	31.62 28	100.00 00
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

Appendix

V

Table of Common Logarithms

From *Rinehart Mathematical Tables*, Rinehart & Company, New York; copyright, 1948 by Harold D. Larsen.

10.0—Four-Place Common Logarithms of Numbers—54.9

N	0	1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396
N	0	1	2	3	4	5	6	7	8	9

55.0—Four-Place Common Logarithms of Numbers—99.9

N	0	1	2	3	4	5	6	7	8	9
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818
96	9823	9827	9832	9836	9841	9845	9850	9854	9869	9863
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996
N	0	1	2	3	4	5	6	7	8	9

Appendix

VI

Bibliography

REFERENCE BOOKS

- Arkin, Herbert, and Colton, Raymond R., *Tables for Statisticians*, Barnes and Noble, Inc., 1950.
- Hauser, Philip M., and Leonard, William R., *Government Statistics for Business Use*, Second Edition, John Wiley and Sons, Inc., 1956.
- Kurtz, Albert K., and Edgerton, Harold A., *Statistical Dictionary of Terms and Symbols*, John Wiley and Sons, Inc., 1939.

MATHEMATICS FOR STATISTICIANS

- Crum, W. L., and Schumpeter, Joseph A., *Rudimentary Mathematics for Economists and Statisticians*, McGraw-Hill Book Co., Inc., 1946.
- Walker, Helen M., *Mathematics Essential for Elementary Statistics*, Henry Holt and Co. Inc., revised edition 1951.

PHILOSOPHY, LOGIC, AND STATISTICS

- Beardsley, Monroe C., *Practical Logic*, Prentice-Hall, Inc., 1950.
- Black, Max, *Critical Thinking*, Prentice-Hall, Inc., 1946.

- Cohen, Morris R., "The Statistical View of Nature," *Journal of the American Statistical Association*, June 1936.
- Cohen, Morris R., and Nagel, Ernest, *An Introduction to Logic and Scientific Method*, Harcourt, Brace and Co., 1934.
- Kline, Morris, *Mathematics in Western Culture*, Oxford University Press, 1953.
- Larrabee, Harold A., *Reliable Knowledge*, Houghton Mifflin Co., 1945.
- Levinson, Horace C., *The Science of Chance*, Rinehart and Co., 1950.

GENERAL ARTICLES

- Cochran, William G., "Research Techniques in the Study of Human Beings," *Milbank Memorial Fund Quarterly*, vol. 33, No. 2, April 1955.
- Cohen, Jerome B., "The Misuse of Statistics," *Journal of the American Statistical Association*, vol. 33, No. 204, December 1938.
- Kuznets, Simon, "Conditions of Statistical Research," *Journal of the American Statistical Association*, vol. 45, No. 249, March 1950.
- Reed, Lowell J., "Man as a Planning Animal," *Journal of the American Statistical Association*, vol. 47, No. 257, March 1952.
- Walker, Helen M., "Statistical Literacy in the Social Sciences," *The American Statistician*, vol. 5, No. 1, February 1951.
- Wilks, S. S., "Undergraduate Statistical Education," *Journal of the American Statistical Association*, vol. 46, No. 253, March 1951.

BOOKS AND ARTICLES ON SPECIAL TOPICS

Graphs and Tables

- Arkin, Herbert, and Colton, Raymond R., *Graphs*, revised edition, Harper and Brothers, 1940.
- Jenkinson, Bruce L., *Bureau of the Census Manual of Tabular Presentation*, United States Government Printing Office, 1949.
- Modley, Rudolph, *How to Use Pictorial Statistics*, Harper and Brothers, 1937.
- Myers, John H., *Statistical Presentation*, Littlefield, Adams & Co., 1956.

Schmid, Calvin F., *Handbook of Graphic Presentation*, The Ronald Press Company, 1954.

Spear, Mary E., *Charting Statistics*, McGraw-Hill Book Company, Inc., 1952.

Quality Control

Duncan, Acheson J., *Quality Control and Industrial Statistics*, Richard D. Irwin, Inc., 1953.

Feigenbaum, A. V., *Quality Control—Principles, Practice, and Administration*, McGraw-Hill Book Co., Inc., 1951.

Fortune Magazine, "Statistical Quality Control," January 1950.

Grant, Eugene L., *Statistical Quality Control*, McGraw-Hill Book Co., Inc., 1952.

Survey Techniques

Blankenship, Albert B., *Consumer and Opinion Research, The Questionnaire Technique*, Harper and Brothers, 1943.

Goode, W., and Hatt, P., *Methods in Social Research*, McGraw-Hill Book Company, Inc., 1952.

Jahoda, Marie, Deutsch, Morton, and Cook, Stuart W., *Research Methods in Social Relations*, Part Two, The Dryden Press, 1951.

Journal of Marketing, "Questionnaire Preparation and Interviewer Technique," October 1949.

Madge, John, *The Tools of Social Science*, Longmans, Green and Co., 1953.

Parten, Mildred B., *Surveys, Polls and Samples*, Harper and Brothers, 1950.

Payne, Stanley L., *The Art of Asking Questions*, Princeton University Press, 1951.

Miscellaneous

Barnes, Leo, *Handbook for Business Forecasting*, 2nd revised edition, Prentice-Hall, Inc., 1949.

Bean, Louis H., "A Simplified Method of Graphic Curvilinear Correlation," *Journal of the American Statistical Association*, December 1929.

Eastwood, R. Parker, *Sales Control by Quantitative Methods*, Columbia University Press, 1940.

- Educational Requirements for Employment of Statisticians*, VA Pamphlet 7-8.9, U. S. Government Printing Office, 1955.
- Ezekiel, Mordecai, *Methods of Correlation Analysis*, 2nd edition, John Wiley and Sons, Inc., 1945.
- Ferber, Robert, *Statistical Techniques in Market Research*, McGraw-Hill Book Co., Inc., 1949.
- Huff, Darrell, and Geis, Irving, *How to Lie with Statistics*, W. W. Norton and Co., Inc., 1954.
- Judd, Clarence, "Do Your Own Forecasting," *Steelways*, January 1950.
- Lorie, James H., and Roberts, Harry V., *Basic Methods of Marketing Research*, McGraw-Hill Book Co., Inc., 1951.
- Mitchell, Wesley C., *The Making and Using of Index Numbers*, Bulletin 656, Bureau of Labor Statistics, United States Government Printing Office, 1938.
- Moroney, M. J., *Facts from Figures*, third and revised edition, Penguin Books, 1956.
- Mudgett, Bruce D., *Index Numbers*, John Wiley and Sons, Inc., 1951.

BOOKS ON SAMPLING THEORY AND PRACTICE

- Cochran, William G., *Sampling Techniques*, John Wiley and Sons, Inc., 1953.
- Deming, William E., *Some Theory of Sampling*, John Wiley and Sons, Inc., 1950.
- Hansen, Morris H., Hurwitz, William N., and Madow, William G., *Sample Survey Methods and Theory*, vol. 1, John Wiley and Sons, Inc., 1953.
- Jastram, Roy W., *Elements of Statistical Inference*, California Book Co., Ltd., 1947.
- Vance, Lawrence L., and Neter, John, *Statistical Sampling for Auditors and Accountants*. New York, John Wiley & Sons, 1956.
- Walker, Helen M., and Lev, Joseph, *Statistical Inference*, Henry Holt and Co., 1953.
- Yates, Frank, *Sampling Methods for Censuses and Surveys*, Hafner Publishing Co., 1953.

GENERAL TEXTBOOKS

- Arkin, Herbert, and Colton, Raymond R., *Statistical Methods*, 4th edition revised, Barnes and Noble, Inc., 1956.
- Blair, Morris M., *Elementary Statistics*, revised edition, Henry Holt and Co., 1952.
- Chaddock, Robert E., *Principles and Methods of Statistics*, Houghton Mifflin Co., 1925.
- Croxton, Frederick E., and Cowden, Dudley J., *Applied General Statistics*, Prentice-Hall, Inc., 1955.
- Freund, John E., *Modern Elementary Statistics*, Prentice-Hall, Inc., 1952.
- Rosander, A. C., *Elementary Principles of Statistics*, D. Van Nostrand Co., Inc., 1951.
- Smith, James G., and Duncan, Acheson J., *Elementary Statistics and Applications*, McGraw-Hill Book Co., Inc., 1944.
- Snedecor, G. W., *Statistical Methods*, 5th edition, Collegiate Press, Inc., of Iowa State College, 1956.
- Walker, Helen M., *Elementary Statistical Methods*, Henry Holt and Co., 1944.
- Wallis, W. Allen, and Roberts, Harry V., *Statistics: A New Approach*, The Free Press, 1956.
- Waugh, Albert E., *Elements of Statistical Method*, 3rd edition, McGraw-Hill Book Co., Inc., 1952.
- Yule, G. Udny, and Kendall, M. G., *An Introduction to the Theory of Statistics*, 14th edition, Charles Griffin and Co., Ltd., 1950.

TEXTBOOKS IN ECONOMICS AND BUSINESS STATISTICS

- Allen, R. G. D., *Statistics for Economists*, Hutchinson's University Library, 1949.
- Croxton, Frederick E., and Cowden, Dudley J., *Practical Business Statistics*, revised edition, Prentice-Hall, Inc., 1948.
- Davies, George R., and Yoder, Dale, *Business Statistics*, John Wiley and Sons, Inc., 1941.
- Kafka, Fritz, *Statistics without Numbers, A Visual Explanation*, Lifetime Editions, Inc., 1950.

- Lewis, Edward E., *Methods of Statistical Analysis in Economics and Business*, Houghton Mifflin Company, 1953.
- Maverick, Lewis A., *Economic Statistics*, Photo-Offset Printing, Carbondale, Illinois, 1949.
- Mills, Frederick C., *Statistical Methods*, 3rd edition, Henry Holt and Co., 1955.
- Neiswanger, William A., *Elementary Statistical Methods*, revised edition, The Macmillan Company, 1956.
- Neter, John, and Wassermann, William, *Fundamental Statistics for Business and Economics*, Allyn and Bacon, 1956.
- Paden, Donald W., and Lindquist, E. F., *Statistics for Economics and Business*, 2nd edition, McGraw-Hill Book Co., Inc., 1956.
- Riggelman, John R., and Frisbee, Ira N., *Business Statistics*, 3rd edition, McGraw-Hill Book Company, Inc., 1951.
- Spurr, William A., Kellogg, Lester S., and Smith, John A., *Business and Economic Statistics*, Richard D. Irwin, Inc., 1954.
- Stockton, John R., *An Introduction to Business Statistics*, D. C. Heath and Co., 1947.

TEXTBOOKS IN OTHER FIELDS OF STATISTICAL APPLICATION

- Hagood, Margaret J., and Price, Daniel O., *Statistics for Sociologists*, Henry Holt and Co., 1952.
- Lindquist, E. F., *A First Book in Statistics, Their Use and Interpretation in Educational Psychology*, Houghton Mifflin Co., 1938.
- McCormick, Thomas C., *Elementary Social Statistics*, McGraw-Hill Book Co., Inc., 1941.
- Odell, C. W., *An Introduction to Educational Statistics*, Prentice-Hall, Inc., 1946.
- Smith, G. Milton, *A Simplified Guide to Statistics for Psychology and Education*, Rinehart and Co., Inc., 1946.

Index

- a , symbol, 241
- Abbreviations, in a table, 37
- Absolute numbers, percentages vs., 472-474
- Absolute variation, 174-200
 - measures of, 198-200
- Acceptance sampling, 451-453
- Accounting ratios, 92
- Achenwall, Gottfried, 6
- A.D., symbol, 182
- Aggregate, in index numbers, 308
- American Society for Quality Control, 451
- American Statistical Association, 4
- Analysis of data, 89-377
 - graphic, 18, 156-158
 - misuses in, 472-477
 - steps, 16-18
 - time series, 221-234
 - for trend, 235-271
- Anderson, Oskar N., cited, 10 n
- Applied statistics, 8-10
- Approximate numbers, 486-489
- Area diagrams, 67-68
- Area sampling, 438
- Arithmetic line graphs, 42-51
 - schematic diagram, 49
- Arithmetic mean, *see* Mean
- Array charts, 115-117
- Arrays, 96-98
 - frequency, 97-98
 - median in, 144-145
 - simple, 96
- Association of American Railroads, 346
- Attributes, sampling by, 452-453
- Average deviation, 174, 181, 182-186
 - coefficient of, 202
 - comparisons, 198-200
 - defined, 182
 - grouped data, 183
 - ungrouped data, 182
- Averages, 127-142, 143-161, 173
 - A.D. and, 186
 - algebraic properties of, 199-200
 - appropriateness of, 170-171, 403-404
 - comparison of the principal, 163-171
 - in computation, 169
 - concept of, 128-129
 - confusion of, 478
 - data arrangement and, 169
 - dynamic, 243
 - effect of extreme values on, 167
 - effect of open-end classes on, 168
 - kinds of, 128
 - location of, 163
 - methods of obtaining, 167
 - moving, 256-261, 274-286
 - position, 143
 - variation and, 198-200
 - varying class intervals and, 168
- Averaging, for index numbers, 308
- Axes, 43
- b , symbol, 241
- Bar charts, 65-68
 - component, 76-78
- Barnes, Leo, 298 n
- Bars, symbol, 129, 150
- Base, index numbers, 306-307
- Base-year weighted aggregate of quantities, 326
- Base-year weighted average of quantity relatives, 326
- Bernoulli, D., 6
- Best fit, line of, 243
- Bias, 432, 467-468
- Bibliography, 508-513
- Bimodal distributions, 123
- Bimodality, 156
- Body, of a table, 36
- Bowley, Arthur L., 203-207
- Box head, of a table, 35
- Business, statistics in, 5-6
- Business Activity, Index of, 305, 327, 346
- C, symbol, 222-223
- Capt, J. C., cited, 38
- Caption, of a table, 35
- Card universe, 432
- Case, 96
- Causation:
 - correlation and, 350
 - covariation and, 480
 - misinterpreted, 479
- Censuses, statistics in, 6
- Centering, 258-260
- Central tendency, measures of, 127
- Chaddock, Robert E., 8 n, 466 n
- Chance, significance and, 420-429
- Chart, defined, 41
- China, statistics in, 6
- Christmas, 273
- Chunk, 447-448
- Clark, John R., 466 n
- Classes:
 - in frequency distribution, 98
 - number of, 104-106
 - open-end, 107-108, 168
- Class frequencies, in a histogram, 117
- Classification:
 - changes in, 466-467
 - incomplete, 468-469
- Class intervals, 105, 107-110
 - averages and, 168
 - defined, 99
 - in graphs, 117
 - varying, 108-110
- Class limits, 106-108
- Cluster sampling, 437-439
 - statistical error, 439
- Coefficient, 174
- Coefficient of average deviation, 202

- Coefficient of correlation, 361-363
 - computation of, 368-369
 - misinterpretation of, 480
- Coefficient of determination, 364-367
- Coefficient of nondetermination, 366-367
- Coefficient of quartile deviation, 175 n, 202
- Coefficient of skewness, 203-205
- Coefficient of variation, 201
- Cohen, Jerome B., 466 n
- Cohen, Morris R., 466 n
- Collection of data, 14-16, 26-30
 - Consumer Price Index, 338-340
 - misuses in the, 466-468
 - Wholesale Price Index, 332
- Comparisons, in semilog graphs, 55
- Component line graph, 79-82
- Component-part graphs, 75-84
- Composite indexes, 303
- Concomitant variation, 349
- Confidence intervals, 415-420
 - sample size and, 443
 - small samples and, 419-420
- Confidence limits, 415-420
 - sample size and, 443
- Constants, 241
- Consumer Price Index, 329, 334-344
- Control charts, 450-460
 - sampling distribution and, 456
 - types of, 457
- Control limits, 456, 458
- Coordinate system, 43
- Corn-hog ratio, 93
- Correlation, 17, 348-377
 - causation and, 350
 - coefficient of, 361-363, 368-369, 480
 - concept of, 348
 - curvilinear, 373-374
 - linear, 373
 - misuse of, 476-477
 - multiple, 374
 - partial, 374
 - rank, 375-377
 - scatter and, 352-355
 - scope of, 373-374
 - simple, 373
- Covariation, 350
 - causation and, 480
- Crude seasonal index, 280
- Current indexes, 329-347
- Current ratio, 92
- Curves:
 - bell-shaped, 121
 - bimodal, 123
 - normal, *see* Normal curve
 - types of, 121, 203
- Curvilinear trend, 262-263, 267-271
- Cycles, 288-296
 - defined, 51, 288
 - measuring, 291-296
 - in semilog graphs, 51-53
- Cyclical variation, 228-229
 - adjustment for, 275
- Δ , symbol, 154
- d , symbol, 137, 191
- D , symbol, 377, 406
- D , symbol, 408
- Data:
 - analysis of, *see* Analysis of data
 - arrangement and the averages, 169
 - classification of, 16-17
 - collection, *see* Collection of data
 - compiling the, 27-30
 - condensation of, 17
 - editing, 27-30
 - homogeneous, 95
 - incomparable, 466
 - index numbers and, 306
 - interpretation, *see* Interpretation of data
 - plotting the, 46
 - presentation, *see* Presentation of data
 - raw, 95
 - as raw material, 13-21
 - rounded, 106, 486-489
 - sources of, *see* Sources of data
 - summarization of, 17
- Deciles, 151, 158
- Deflating, defined, 232
- Δ , symbol, 154
- Deming, William Edwards, 447 n
- Democracy, statistics and, 3-4
- Department Store Sales, Index of, 305, 345-346
- Dependent variable:
 - defined, 42
 - in trend, 241
- Descriptive statistics, defined, 19
- Determination, coefficient of, 364-367
- Deviations, from the mean, 140-142
- Difference, 406-409
 - chance, 420
 - real, 420
 - standard error of a, 406-409
 - true, 425-426
- Direct investigation, 25
- Dispersion:
 - in control charts, 457
 - measures of, 127, 173-202
 - neglect of, 475
- Disproportionate sampling, 435
- Distributions, in sampling, 389-398
- Ditto marks, in a table, 38
- Duncan, Acheson J., 458 n
- Dutka, Sol, 433 n, 438 n
- Easter, 273
 - index adjustment for, 346
- Economics, statistics in, 5-6
- Editing, of data, 27-30
- Education, statistics and, 3-4
- Educational Requirements for Employment of Statisticians*, V A Pamphlet, 7 n
- Egypt, statistics in, 6
- Elementary sampling unit, 437
- Engelberg, Carl C., cited, 5 n
- England, statistics in, 6
- Entry form, 99-101
- Enumeration, incomplete, 468
- Estimating equation, 356-357
- Estimating line, 356-357

- Estimation, 413-420
 concept, 387-389
 sample size and, 419-420
 Extrapolation, 298, 357 n
 faulty, 476
 Extreme values:
 effect on averages, 167
 variation and, 198-199
- f*, symbol, 135
 Federal Reserve Board, indexes, 305, 345
 Feigenbaum, A. V., 450 n, 453 n
 Ferber, Robert, 448 n
 Field, of a table, 36
 Fisher, R. A., statistical pioneering, 7
 Fluctuations, comparison of, 60
 Footnotes, 36, 38, 47
 Forecasting, 254 n, 297-299
 Frankel, Lester R., 433 n, 438 n
 Freehand method, for trend, 237
 Frequencies:
 absolute, 110
 cumulative, 111-114, 123-126
 defined, 97, 99
 percentage, 110-111
 relative, 110
 sequence of, 105
 total, 102
 Frequency array, 97-98
 Frequency densities, 109
 Frequency distribution, 95-114
 constructing a, 104-110
 cumulative, 112, 123
 graphic presentation, 115-126
 location of the averages in, 163
 in sampling, 389-398
 skewed, 121-123, 164-167
 symmetrical, 163-164
 see also Universe distribution, Sample distribution, and Sampling distribution
 Frequency graphs, 115-126
 Frequency polygon, 119-123
 area under, 119
 shapes of, 121-123
 Frequency table, 101
- G*, symbol, 160
 Galton, Sir Francis, 356
 statistical pioneering, 7
 Gauss, C. F., 6
 Geis, Irving, 480 n
 Geometric forms, 65-69
 Geometric mean, 159-161
 Geometric progression, averaging a, 161
Government Statistics for Business Use,
 Hauser and Leonard, 32
 Grant, Eugene L., 452 n, 458 n
 Graphic analysis, 156-158
 Graphs:
 averages from, 170
 combination, 73-75
 defined, 41
 formal requirements, 84-85
 frequency, 115-126
 misuses, 469-473
 tables compared with, 85
 technical aspects, 48-50
 time series, 233
 see also Arithmetic line graphs, Bar graphs, etc.
 Greece, statistics in, 6
 Grid, defined, 42
 Grid lines, 43
 Guessed mean, 188
- Hansen, Morris H., 411 n, 433 n
 Hauser, Philip M., cited, 32
 Headings, in a table, 38
 Headnotes, 36, 47
 Hebrews, statistics, 6
 Herodotus, 6
 Heterogeneity, in a sampling, 437
 Histogram, 117-120
 area under, 119
 construction, 118
 mode and, 158-159, 170
 Historical thinking, 10-11
 Homogeneity, in sampling, 434
 Horizontal axis, defined, 42
 Huff, Darrell, 480 n
 Hurwitz, William N., 411 n, 433 n
- i*, symbol, 139
J, symbol, 222-223
 Independent variable:
 defined, 42
 in trend, 241
 Indexes:
 composite, 303
 defined, 302 n
 faulty use of, 476
 Index numbers, 302-329
 base, 306-307
 classification of, 305
 combining, 323
 in computation, 322 n
 current, 329-347
 data and, 306
 defined, 302
 making comparable, 321-323
 percentage change in, 324
 sample in, 304
 special problems, 321
 splicing, 323-324
 weighting, 313-321
 Indicators, 289
 Induction, 383-384
 Inductive statistics, 19, 381-462
 Industrial Production, Indexes of, 305, 342, 344-345
 Inference, statistical, 413-429
 Inquiry, 26
 Inspection, for determining trend, 237
 Institute for Research, Chicago, Ill., 8
 Interpretation of data, 18-19, 381-462
 misuses in the, 477-480
 Interquartile range, skewness and, 206
 Interval estimate, 415-420
 Irregularity, 296-297
 Irregular variations, adjustment for, 278
 Item, 96

- Jastram, Roy W., 387 n
 J curve, 122
 Jenkinson, Bruce L., 38 n
 Jobs, in statistics, 8-10
 Judd, Clarence, 299 n
 Judgment sample, 445
- k^2 symbol, 366
 Kendall, Maurice G., 11 n, 376 n
 Kline, Morris, 480 n
 Kornhauser, Arthur, 26 n
 Kurtosis, 127, 207-208
 defined, 203
 measures of, 207-208
 Kuznets, Simon, 16 n
- l , symbol, 147
 Labor indexes, 347
 Labor Statistics, Bureau of, 304, 305, 321, 329-344
 indexes, 347
 Labor turnover, 93
 Lankford, Francis G., 466 n
 Larrabee, Harold, 351, 466
 Laspeyres, Etienne, 315 n
 Laspeyres formula, 315, 333
 "Leader work," 34
 Least-squares curves, 256 n
 Least-squares method, for trend, 240-253, 267-271
 Leonard, William R., cited, 32
 Leptokurtic, defined, 203
 Lev, Joseph, 420 n
 Levinson, Horace C., 466 n, 468 n
 Line graph, 79-82
 Littauer, Sebastian B., 451 n
 Logarithm, defined, 50 n
 Logarithmic graph, double, 51 n
 Logarithmic mean, 160
 Logarithmic scale, 51, 52
 Logarithms, table of, 506-507
- μ , symbol, 390
 M.D., symbol, 182
 Madow, William G., 411 n, 433 n
 Magnitude, 96
 Magnitudes, in line graphs, 58
 Mail questionnaires, 26
 Maps, 71-73
 Mass, 95
 Masses, measurement of, 127-220
 Mathematical statistics, 8-10
 Mean, 127-142
 A.D. and, 186
 appropriateness, 170-171, 403-404
 characteristics of, 139-141, 169
 in computation, 169
 in control charts, 457
 data arrangement and, 169
 defined, 129
 deviation properties, 140-142
 effect of extreme values on, 167
 estimation of the, 413-417
 geometric, 159-161
 from grouped data, 134-139
 guessed, 136-139
 location, 163-170
 long method, 134-136
 mathematical properties, 139-142, 169
 of means, 131
 median confused with, 478
 methods of obtaining, 167
 misuse of the, 474
 mode confused with, 478
 modified, 280
 open-end classes and, 168
 as a point of balance, 140
 in sampling, 393-395
 short method, 136
 significance of a difference, 422-425, 425-427
 substitution property, 141-142
 true, defined, 390
 from ungrouped data, 129-130
 of the universe, 390
 varying class intervals and, 168
 weighted, 130-134
 Mean deviation, 182
 Measure, 96
 Measurement errors, 431
 Median, 143-151
 A.D. and, 186
 appropriateness, 170-171, 403-404
 characteristics of, 150, 169
 in computation, 169
 data arrangement and, 170
 deviation property, 150
 difference property, 150
 effect of extreme values on, 167
 faulty use of the, 475
 graphic analysis, 156-158
 for grouped data, 146-149
 location, 163-170
 mathematical properties, 150, 169
 mean confused with, 478
 methods of obtaining, 167
 ogive and, 157, 170
 open-end classes and, 168
 semi-interquartile range and, 179
 standard error of, 402-404
 true, 403, 418
 for ungrouped data, 144
 varying class intervals and, 168
 Mesokurtic, defined, 203
 Mid-point assumption, 103-104, 136
 Mills, Frederick C., 307 n
 Misuses of statistics, 465-480
 Mitchell, W. C., 15 n
 Modal class, 153
 Mode, 151-156
 appropriateness, 170-171
 in computation, 169
 crude, 152
 data arrangement and, 170
 effect of extreme values in, 167
 faulty use of the, 475
 for grouped data, 152
 histogram and, 158-159, 170
 location, 163-170
 mean confused with, 478
 methods of obtaining, 167
 open-end classes and, 168

- Mode (*Continued*)
 for ungrouped data, 152
 varying class intervals and, 168
 Mode, Elmer B., 466 n
 Modified mean, 280
 Modley, Rudolph, 70 n
 Moving averages, 256-261, 274-286
 μ , symbol, 390
 Mudgett, Bruce D., 317 n
 Multiple correlation, 374
 Multistage sampling, 438

 n , symbol, 395, 420
 N , symbol, 129
 Nagel, Ernest, 466 n
 National Bureau of Economic Research,
 cycle study, 290, 299
 National Industrial Conference Board, 321
 Negative correlation, 354
 Neiswanger, William A., 466 n
 Neurath, Otto, 70 n
New York Times, indexes, 305, 327, 346
 Nondetermination, coefficient of, 366-367
 Normal, statistical, 291
 Normal curve, 121, 193-198
 absolute variation in, 199
 areas under, 213, 214-220
 defined, 193-195, 220
 ordinates and, 210-214
 practical uses of, 210-220
 standard deviation and, 195-198
 Normal equations, 244
 Null hypothesis, 421-422

 O , symbol, 222-223
 Observation, 25, 96
 Observation equations, 244
 Office of Statistical Standards, 31
 Ogive, 117, 123-126
 median and, 157, 170
 Open-end classes, 107-108, 168
 Optimum allocation, 435
 Origin:
 in line graphs, 42
 shifting the, 254, 267
 trend and, 240, 254
 zero, 43-44
 Overlay scale, 62-63

 $\pm$, symbol, 179 n
 p , symbol, 232, 309, 404, 457
 P , symbol, 309, 405
 Paasche, Hermann, 315 n
 Parabolas, trend, 263
 Parameters, 390
 Partial correlation, 374
 Partition, measure of, 181
 Peakedness:
 measures of, 207-208
 see also Kurtosis
 Pearson, Karl, 165-167
 coefficient of correlation, 362-364
 measure of skewness, 203-205
 statistical pioneering, 7
 variation coefficient, 201
 Percentage change, on semilog graph, 60-63
 Percentage frequencies, cumulative, 125
 Percentages:
 absolute numbers vs., 472-474
 cautions concerning, 91-92
 faulty use of, 472-474
 as ratios, 90-92
 Percentiles, 151, 158
 Percents, averaging, 133
 Perfect correlation, 354
 Periodic movements, 224-227
 Personal interview, 26
 Pictographs, 69-71
 component, 78-79
 Pie diagram, 80
 Platykurtic, defined, 203
 Plotting:
 arithmetic line graphs, 46-47
 on a logarithmic scale, 53
 Point estimate, 414
 Population, defined, 383
 Population changes, 233
 Positional measures, 151
 graphic analysis, 156
 variation, 174-181
 Positive correlation, 353
 Predicted variable, 357
 Predicting variable, 357
 Prefatory note, of a table, 36
 Prejudice, 19-20
 Presentation of data, 16
 combination graphs, 73-75
 component-part, 75-84
 geometric forms, 65-85
 informal, 33
 line graphs, 41-64
 maps, 71-73
 misuse in, 468-477
 pictures, 69-71
 semitabular, 34
 tables, 33-40
 textual, 33-34
 Price changes, adjustment for, 232
 Price indexes, 306-325
 current, 330-341
 Price relative, defined, 309-310
 Primary sampling unit, 437
 Primary source, defined, 30
 Probability sampling, 431-445, 446-447
 Problems, data and, 13-15
 Process control, 453-454
 Processing errors, 431
 Production-Worker Employment, Index of,
 347
 Product-moment approach, 367
 Proportion:
 significance of a difference, 425, 427
 standard error of a, 404-406
 true, 418
 Proportionate sampling, 435
 Punch card, 28, 29
 Purposive sampling, 445-447
 random sampling compared with, 446-447

 q , symbol, 232, 313, 406
 Q , symbol, 176, 325, 405
 Q_1 , Q_3 , symbol, 151

- Quadrants, 43
 - in line graphs, 42
- Quality control, statistical, 450-460
- Quantity indexes, 325-326
 - current, 341-345
- Quantity weights, 313
- Quartile deviation, 174, 175-181, 201 n
 - coefficient of, 202
- Quartiles:
 - defined, 151
 - graphic analysis, 158
 - grouped data, 177-179
 - skewness and, 205
 - ungrouped data, 176-177
- Quételet, L. A. J.; 6
- ρ , symbol, 376
- r , symbol, 368, 376 n
- r^2 , symbol, 367
- Railroad Operating Efficiency, Index of, 346
- Rameses II, census, 6
- Random numbers, 432, 461-462
- Random samples:
 - defined, 432
 - restricted, 433-434
 - unrestricted, 433
- Random sampling, 431-445
 - purposive sampling and, 446-447
 - unrestricted, 439
- Range, 174
 - in control charts, 457
 - crude, 174-175, 181-182
 - faulty use of the, 475
- Rank correlation, 375-377
- Rare, defined, 424
- Rare sample, 400-401
- Rates, averaging, 161
 - defined, 90
 - in semilog graphs, 55-58
- Ratio chart, semilog graph as, 54-55
- Ratios, 89-93
 - averaging, 160
 - defined, 89
 - to moving average, method of, 274-286
 - in semilog charts, 60-62
 - to trend, method of, 274-286
 - types of, 90-91
- Reed, Lowell J., 7 n
- Regression equation, 357
 - see also* Trend equation
- Regression line, 356-357
 - computation of, 370-373
- Relative, defined, 90
- Relative variation, 200-202
- Report, statistical, 483-485
- Residual method, 291-296
- Response errors, 431
- Restricted random samples, 433-434
- Rho , symbol, 376
- Rome, statistics in, 6
- Rounding, 106, 486-489
- Ruling, in a table, 38
- σ , symbol, 187, 365, 390
- Σ , symbol, 129
- S , symbol, 222-223
- s_y , symbol, 361
- S_r , symbol, 359, 361
- Sample:
 - biased, 467-468
 - defined, 383
 - in index numbers, 304
 - random, 390 n
 - rare, 400-401
 - size, *see* Sample size
 - small, 410-411, 419-420
 - universe and, 384-389
- Sample distribution, 390-391, 394
- Sample mean:
 - common, 423-425
 - universe mean and, 422-425
 - unusual, 423-425
- Sample proportion, universe proportion and, 425
- Sample size, 410-411, 441-445
 - confidence interval and, 443
 - confidence limits and, 443
 - estimation and, 419-420
 - in purposive sampling, 446
 - sequential sampling and, 449
 - standard deviation and, 442
 - standard error and, 410-411
 - standard error of the mean and, 441-442
 - standard error of the proportion and, 441-442
- Sampling, 381-462
 - acceptance, 451-453
 - cluster, 437-439
 - practice, 430-449
 - probability, 431-445
 - purposive, 445-447
 - random, 431-445
 - reason for, 385-386
 - sequential, 448-449, 452
 - systematic, 439-441
 - types of, 447-449
- Sampling designs, 430-445
- Sampling distribution, 391-393, 394
 - control chart and, 456
- Sampling error, 398, 431
- Sampling interval, 440
- Sampling start, 440
- Sampling tables, 450-460
- Sampling theory, 381-412, 413-429
- Sampling units, kinds of, 437
- Scale, in a semilog graph, 51
- Scale breaks, 44-45
- Scale caption, 47, 54
- Scales, 43
 - in line graphs, 42, 45-46
 - in semilog graphs, 55-56
- Scatter, 173
 - correlation and, 352-355
- Scatter diagram, 352, 355
- Schorling, Raleigh, 466 n
- Seasonal index, 280
- Seasonal relative, 280
- Seasonal variations, 224-228, 272-287
 - adjustment for, 286-287
- Secondary source, defined, 30
- Selected-points method, for trend, 238
- Semi-average, defined, 238 n

- Semi-average method, for trend, 238-240
 Semi-interquartile range, 174-181, 201 n
 comparisons, 198-200
 median and, 179
 skewness and, 206
 Semilogarithmic line graphs, 50-63
 construction of, 51-54
 limitations of, 63
 Semilog trend, 262
 Semitabular presentation, 34
 Sequential sampling, 448-449, 452
 sample size and, 449
 Shewhart, Walter A., 451
Sigma, symbol, 129, 187, 365, 390
 Significance:
 limitations of tests of, 428
 statistical, 420-429
 Significance of difference, 387-389
 Simple aggregate of actual prices, 308-309
 shortcomings of, 311-313
 Simple aggregate of quantities, 325
 Simple average of price relatives, 309-310
 shortcomings of, 311-313
 Simple average of quantity relatives, 326
Sk, symbol, 203
 Skewness, 121-123
 defined, 203
 measurement of, 127, 202-207
 quartiles and, 205
 Slopes, in semilog graphs, 56
 Small sample, 419-420
 standard error and, 410-411
 Smoothing, 119 n
 Snedecor, George W., cited, 3 n
 Source notes, 36, 47, 54
 Sources of data, 30-32
 Spearman, Charles Edward, 376-377
 Special-purpose indexes, 305, 327, 346-347
 Specific seasonal, 273
 Splicing, index numbers, 323-324
 Spurious accuracy, 469-470
 Spurious correlation, 351
 Squared deviations, from the mean, 141
 Square roots:
 computation, 490-492
 table of, 494-503
 Squares, table of, 494-503
 Standard deviation, 174, 181, 186-193
 comparisons, 198-200
 defined, 186-187
 grouped data, 190-193
 normal curve and, 195-198
 in sampling, 395
 sample size and, 442
 standard error of, 402-404
 true, 403, 418
 ungrouped data, 187-189
 use of, 197-198
 see also Standard error, and Standard error of specific measure
 Standard error, 395
 cluster sampling, 439
 sample size and, 410-411
 small, 430
 universe size and, 411
 Standard error of a difference, 406-409
 Standard error of estimate, 357-360
 computation of, 369-370
 see also Standard deviation
 Standard error of the mean, 393, 395
 in estimation, 414-417
 formula for, 396
 interpretation of, 398-402
 sample size and, 441-442
 Standard error of the median, 402-404
 Standard error of a proportion, 404-406
 sample size and, 441-442
 Standard error of the standard deviation, 402-404
 Standard error of the total, 404
 Statistic, defined, 391
Statistical Abstract of the United States, 31
 Statistical administrator, 8-10
 Statistical analysis, 89-377
 Statistical assistant, 8-10
 Statistical normal, 291
 Statistical report, 483-485
Statistical Services of the United States Government, 32
 Statistical thinking, 10-11
 Statistical work, types of, 8-10
 Statistician, qualifications of a, 7-11
 Statistics:
 defined, 4-7
 function of, 1-21
 history of, 6-7
 importance of, 3-11
 Step deviations, 137
 Stratification, 434-437
 Stratified sampling, 434
 Stub, of a table, 35, 37
 Symbols, statistical work, 129 n
τ, symbol, 376 n
T, symbol, 222-223
 Tables, 33-40
 construction of, 37-38
 defined, 34
 frequency, 101
 graphs compared with, 85
 how to read, 38-40
 parts of, 35-37
 types of, 34-35
Tabular Presentation, Bureau of the Census Manual of, Jenkinson, 38 n
 Tabulation, mechanical, 27
 Tally sheet, 29, 99-101
Tau, symbol, 376 n
 Time series, 221-234
 graphs, 233
 index numbers as, 304
 Title:
 in a graph, 47, 54
 of a table, 35, 37
 Total, standard error of a, 404
 Totalling, for index numbers, 308
 Trend, 223-224, 235-271
 adjustment for, 261-262, 275
 basic equation for, 244
 curvilinear, 262-263, 267-271
 defined, 223
 eliminating, 236

- extrapolation of, 476
- forecasting, 298
- measurement of, 237-253
- misinterpretation, 479
- origin and, 240, 254
- Trend equation:
 - conversion of, 264-267
 - development of, 244-253
 - use of the, 253-256
 - see also* Regression equation
- True difference, 408
- True mean, 390
- True median, estimate, 418
- True proportion, 405
 - estimate, 418
- True standard deviation, estimate, 418
- True total, estimate, 418
- Truth, 19-20
- t*-table, 419-420
- Typical seasonal, 273
- Typical seasonal relative, 280

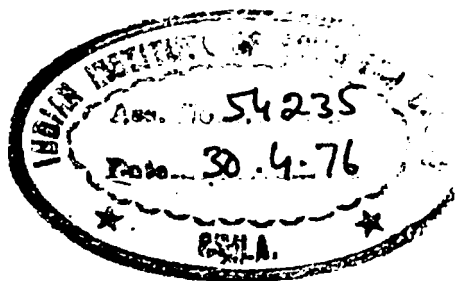
- U curve, 123
- Uniformity, 174
 - Q* and, 181
- United Nations, index, 345
- United States, indexes, *see* name of specific index
- Universe:
 - defined, 383
 - finite, 384, 411
 - infinite, 384, 411
 - sample and, 384-389
 - size and standard error, 411
- Universe distribution, 390-391, 394
- Universe mean, sample mean and, 422-425
- Universe proportion, sample proportion and, 425
- Unusual, defined, 424

- v*, symbol, 232
- V*, symbol, 201-202, 326
- Value, 96
- Value indexes, 305, 326, 345-346
- Value weights, 313
- Variables:
 - continuous, 102
 - defined, 102
 - discontinuous, 102
 - discrete, 102
 - kinds of, 42
 - sampling by, 453
- Variance, 365-366
 - defined, 187
- Variate, 96
- Variation:
 - coefficient of, 201
 - computed measures, 181-193
 - concomitant, 349
 - measures of, 127, 173-202
 - positional measures, 174-181
 - Q* and, 181
 - see also* Absolute variation, Relative variation, Average deviation, Standard deviation, and Standard error
- Variations:
 - calendar, 231
 - cyclical, 228-229, 288-296
 - irregular, 229-230, 296-297
 - misinterpretation, 479
 - seasonal, 224-228, 272-287
- Vertical axis, defined, 42
- Volume diagrams, 68-69

- Walker, Helen M., 4 n, 420 n
- Ward, Barbara, 288 n
- Weighted aggregate of actual prices, 314
- Weighted average of price relatives, 317
- Weighting, index numbers, 313-321
- Weights:
 - defined, 130-131
 - given-year, 315
 - typical-year, 317
- Wells, H. G., cited, 3
- Wholesale Price Index, 304, 305, 324, 330-334
- Wilcox, Walter F., 7 n
- Wilks, S. S., cited, 3 n
- World Index of Industrial Production, 345
- Wright, Carroll D., 4
 - cited, 20

- x*, symbol, 137
- |x|*, symbol, 150
- X*, symbol, 129
- x̄*, symbol, 129
- X̄*, symbol, 456 n
- X*-axis, defined, 42

- y*, symbol, 362
- Y*, symbol, 242 n
- Y_e*, symbol, 242 n
- Y*-axis, defined, 42



OTHER BOOKS OF INTEREST

A New Text Book of

Analytical Solid Geometry

S. M. Mathur

VOLUME I

This book is designed as an introductory textbook for students appearing at B. Sc., B.E. and M.Sc. examinations of all Indian and overseas universities. The thorough study of this volume will help the student to proceed further to the new frontiers of knowledge, carrying a fundamental background of the classical Analytical Solid Geometry.

Rs. 10.00

A New Text Book of

Analytical Solid Geometry

S. M. Mathur

VOLUME II

This volume is designed as a continuous chain of study following the volume I. This volume will serve the purpose of the students appearing for B.A./B.Sc./B.E. and M.Sc. examinations of all Indian and foreign universities. This volume is meant to open further frontiers of knowledge for the students and to give an understanding and the preparation which may help them to proceed further to acquire the present-day knowledge of some basic concepts of Differential Geometry and Theory of Surfaces.

Rs. 12.50

A Text Book of Analytical Geometry & Vector Analysis

R. M. Khan

This treatise deals with analytical geometry (two and three dimensions) and vector analysis in a very simple way from elementary considerations to rigorous treatments. It gives the ideas of straight lines, planes, different curves with their preparation in cartesian and polar coordinates, different quadrics with usual properties and elementary vectors needful for various branches of science in a very simple way. It contains numerous diagrams and suitable examples meant for general as well as technical students.



Library

IIAS, Shimla



00054235

OXFORD & IBH PUBLISHING
Calcutta :: Bombay :: New Delhi